# Eliciting Expert Uncertainty from Decision Making

**A thesis**

**submitted in partial fulfilment**

**of the requirements for the Degree**

**of**

**Doctor of Philosophy in Statistics**

**at**

**The University of Waikato**

by

**J. R. Falconer**


THE UNIVERSITY OF
WAIKATO
*Te Whare Wānanga o Waikato*

**2024**

# Abstract

Eliciting expert uncertainty is a complex task with many caveats. Its usefulness in high-level decision-making and Bayesian inference makes it a necessary task that must be completed accurately. This thesis explores uncertainty elicitation in terms of prior elicitation for Bayesian inference and the wider knowledge elicitation field. It describes methods currently available for prior elicitation and proposes a new typology, highlighting lesser-known methods. Based on the limitations of current methods, a new method for prior elicitation is introduced. This new method allows an analyst to model expert decision-making data to elicit a probability distribution that reflects expert uncertainty. An example of parole board decision-making is used to elicit a prior distribution of a prisoner re-offending upon release from prison. This example shows analysts how to elicit distributions for tabular data; however, more complex data types, such as images and reports, are often used for decision-making. To elicit distributions capturing expert uncertainty from more complex data, this thesis introduces a deep learning approach and uses an example of eliciting cancer risk, in histopathology, to illustrate this approach for the wider knowledge elicitation field.

# Note on Publications

This thesis is with publications. The details of the publications are as follows:

**Chapter 3**: Published Online

Julia R Falconer, Eibe Frank, Devon LL Polaschek, and Chaitanya Joshi. Methods for eliciting informative prior distributions: A critical review. *Decision Analysis*, 19(3):189–204, 2022

**Chapter 4**: Accepted for Publication

Julia R Falconer, Eibe Frank, Devon LL Polaschek, and Chaitanya Joshi. Eliciting informative priors by modelling expert decision making. *Decision Analysis*, 2023

**Chapter 5**: Submitted

Julia R Falconer, Eibe Frank, Devon LL Polaschek, and Chaitanya Joshi. Utilizing deep learning to elicit expert uncertainty. *Submitted to The American Statistician*, 2023

# Acknowledgements

Firstly, I want to acknowledge my supervisors, Doctor Chaitanya Joshi, Professor Eibe Frank and Professor Devon Polaschek. Thank you for the countless hours you spent reviewing, editing, and engaging in in-depth discussions with me. This thesis would not have been possible without your support and assistance. I would also like to thank the staff and students in the Department of Mathematics at the University of Waikato who made my time there thoroughly enjoyable and for the endless advice and support when times were tough.

I would like to express my gratitude to Andrea Haines for her help with my writing. From running workshops to reviewing my article submissions, Andrea has been instrumental in helping me become a better writer. I am forever indebted to her for taking the time to support me.

Lastly, I want to say a huge thank you to my support crew, my loving family, who shared in the highs and lows of this experience; especially my husband, Scott, thank you for being my everything.

# List of Figures

# List of Tables

# Contents

# Chapter 1

# Introduction

In the first years of schooling, we teach children to understand basic uncertainty through words, such as certain, possible or impossible. When the human life is intrinsically uncertain, having a good understanding of uncertainty is important and sometimes crucial for critical events [1]. Quantifying uncertainty may help in understanding uncertainty for some real-life applications, but importantly, it also allows for uncertainty to be implemented into computational algorithms. When uncertainty is quantified in the form of a probability distribution an outcome can be assessed to be uncertain or certain by looking at the central tendencies of the distribution (mean, mode, median) and the shape of the distribution; for example when assessing the probability of an outcome, if the central tendencies are around 0.5 and distribution is wide then the outcome can be interpreted as more uncertain than if the central tendencies are closer to zero or one and the distribution is narrower.

Where there is no data available to quantify these distributions, it is common practice to elicit these distributions from expert knowledge (*expert knowledge elicitation*). Experts can be used to elicit both *aleatoric* (uncertainty from random variation that is irreducible [2]) and *epistemic* uncertainty (uncertainty from lack of knowledge that is reducible [2]). Expert knowledge elicitation research tends to focus on eliciting uncertainty through interviewing experts [3].

While this approach has demonstrated its reliability in various scenarios, there are instances where it may not be suitable. Other methods for knowledge elicitation are less known and, hence, not used as frequently in practice. Methods that use experts to elicit probability distributions have many limitations [3]; however, methods that require no expert input are often impractical [4].

This thesis outlines existing methods that can help an analyst elicit probability distributions that capture the uncertainty surrounding an event and introduces a new method that reduces some of the limitations found in other methods. This new method uses information from a related decision-making task to infer a probability distribution.

## 1.1 Thesis Objectives

The main objective of this thesis is to develop a novel approach for quantifying expert uncertainty that addresses the limitations of existing methods. This method is first introduced under the field of prior elicitation and then extended to the wider knowledge elicitation field.

To achieve this, the following research objectives must be implemented:

1. Align the field of prior elicitation by:
    - Giving an accurate overview of the current state of methods of obtaining informative prior distributions.
    - Creating a typology for the different types of methods.
    - Highlighting the drawbacks of current methods and giving precise direction for future research.
    - Exploring the relationship between prior elicitation and decision making.

2. Propose a method that eliminates some of the drawbacks of current methods and helps in addressing potential biases.

3. Provide modern statistical solutions to the proposed method.

## 1.2 Contributions

The main contributions of this research are:

1. Expanding the field of prior elicitation by including all forms of obtaining an informative prior; and creating and formalising definitions for groups of existing methods to facilitate better understanding and communication among researchers.

2. Introducing a new method of prior elicitation, utilising expert decision-making, that eliminates the need for an expert to have statistical knowledge, eliminates the need to interact with experts (reducing bias) and puts an emphasis on obtaining accurate informative prior distributions. Our method also allows analysts to explore variables that may be considered to introduce bias in the decision-making process.

3. Using methods of artificial intelligence to elicit and quantify expert uncertainty, allowing our methods to be used when decision-making is based on more complex data, like reports and images.

## 1.3 List of Publications

During the course of this research, the following articles have been published in and accepted for publication in peer-reviewed journals:

Julia R Falconer, Eibe Frank, Devon LL Polaschek, and Chaitanya Joshi. Methods for eliciting informative prior distributions: A critical review. *Decision Analysis*, 19(3):189–204, 2022

Julia R Falconer, Eibe Frank, Devon LL Polaschek, and Chaitanya Joshi. Eliciting informative priors by modelling expert decision making. *Accepted to Decision Analysis*, 2023

The following articles have been submitted to peer-reviewed journals:

Julia R Falconer, Eibe Frank, Devon LL Polaschek, and Chaitanya Joshi. Utilizing deep learning to elicit expert uncertainty. *Submitted to The American Statistician*, 2023

### 1.3.1 Chapter Outline

The thesis is organised as follows:

**Chapter 2** provides background information that is beneficial to understanding the need for prior elicitation and the key aspects of artificial intelligence that are used in this thesis.

**Chapter 3** sets out to review the field of prior elicitation by reviewing current methods and highlighting their limitations.

**Chapter 4** introduces a new method for prior elicitation.

**Chapter 5** extends the method in Chapter 4 to the field of knowledge elicitation and provides solutions in deep learning to address the limitations discussed in Chapter 4.

Finally, the thesis is concluded in **Chapter 6** and further research paths are introduced.

## References

[1] M Granger Morgan. Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National academy of Sciences*, 111(20):7176–7184, 2014.

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

[3] Anthony O'Hagan, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. Uncertain judgements: eliciting experts' probabilities. 2006.

[4] Julia R Falconer, Eibe Frank, Devon LL Polaschek, and Chaitanya Joshi. Methods for eliciting informative prior distributions: A critical review. *Decision Analysis*, 19(3):189–204, 2022.

[5] Julia R Falconer, Eibe Frank, Devon LL Polaschek, and Chaitanya Joshi. Eliciting informative priors by modelling expert decision making. *Accepted to Decision Analysis*, 2023.

[6] Julia R Falconer, Eibe Frank, Devon LL Polaschek, and Chaitanya Joshi. Utilizing deep learning to elicit expert uncertainty. *Submitted to The American Statistician*, 2023.

# Chapter 2

# Background

This thesis delves into quantifying expert uncertainty through expert knowledge elicitation. It explores methods that can be used for both prior elicitation and future decision-making and introduces methods that utilise Bayesian logistic regression and deep learning models. This chapter provides readers with background information that will aid in comprehending the concepts discussed in upcoming chapters.

## 2.1 Decision Making Applications

When it comes to high-level decision-making, it is important for decision-makers to understand the uncertainty surrounding potential outcomes [1]. A probability distribution quantifying uncertainty can be used as one piece of information for decision-makers to come up with important decisions. When attempting to have rational decision-making, decision-makers should have all the available information to make a rational decision and should be consistent across all their decisions. However, there are a lot of decision-making processes that require decision-makers to make decisions the best they can with limited information and time, such as those made in the criminal justice field. These decision-making processes may be considered bounded rational decision-making because although the decision-maker is aiming for rational decisions they are limited by time and may not have all the information to

make the most rational decision [2]. Some methods for knowledge elicitation are not well suited to these bounded rational decision-making tasks as the elicitation method itself requires a lot of time to form a useful distribution. However, using quantified uncertainty in bounded, rational decision-making is still important and should be considered in research.

In this thesis, the term "expert decision-making" [3, 4] is used to define a decision-making process that is completed by the expert whose uncertainty is to be quantified. This helps in distinguishing between the decision-making process that is used to elicit uncertainty (expert decision-making) and the future/high-level decision-making task for which the uncertainty may be used.

## 2.2 Bayesian Inference

Statistical inference has two main branches: frequentist inference and Bayesian inference. While the frequentist approach uses only the observed data for inference, Bayesian inference allows analysts to use both the data and prior information [5]. There is continual debate over which method is more suitable for specific tasks [6, 7]. This thesis focuses on a unique scenario where there is limited or no data to make predictions. In such situations, the frequentist approach is not appropriate due to the requirement of data for inference. Instead, Bayesian inference is more suitable as it permits the use of prior information to make inferences, even with limited data.

Bayesian inference follows from Bayes' theorem, which is a theorem that gives the conditional probability of $A$ given $B$, also known as the posterior probability of $A$. This theorem utilises the prior probabilities of $A$ and $B$ and the conditional probability of the event $B$ given $A$ (as shown in Equation 2.1).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.1}$$

This can be simplified to the posterior probability of $A$ given $B$ being proportional to the prior probability of $A$ multiplied with the conditional probability of the event $B$ given $A$ (the likelihood) (Equation 2.2).

$$P(A|B) \propto P(B|A)P(A) \tag{2.2}$$

For inference, we wish to obtain posteriors on model parameters, $\theta$, from the data, $D$. Equation 2.2 can be rewritten as

$$P(\theta|D) \propto P(D|\theta)P(\theta). \tag{2.3}$$

When data is limited or nonexistent, the likelihood $P(D|\theta)$ will not have a significant impact on the posterior distribution. Instead, the posterior distribution will be strongly influenced by the prior information, $P(\theta)$. It is common practice in Bayesian inference to use a non-informative prior distribution for $P(\theta)$. However, when we have limited information from the likelihood, the prior must provide information to obtain a reasonable posterior distribution. These types of priors are called *informative* prior distributions.

The posterior distribution for many inference problems is intractable. To complete Bayesian inference for these problems, an analyst will have to find an approximation of the posterior distribution. This thesis refers to two posterior approximation methods: Markov Chain Monte Carlo (MCMC) and variational Bayes. MCMC is a simulation method that is a step-by-step process. At each step, a sample will be drawn from an approximate distribution; then, the sample will be corrected to better align with the target posterior distribution [8]. The goal is that the approximation will converge to the target posterior

distribution. Variational Bayes is a method to approximate the posterior distribution, $P$, by taking a distribution, $Q$, from a family of distributions of a simpler form than $P$ [8]. The goal of variational Bayes is to find a $Q$ that minimises the Kullback-Leibler divergence (Equation 5.1).

$$D_{KL}(Q, P) = \sum Q log(\frac{Q}{P}) \tag{2.4}$$

### 2.2.1 Prior Elicitation

Obtaining a precise and informative prior distribution can be a daunting task, especially when time is a constraint. However, accurate results are achievable with the right approach and methodology. One common method for obtaining informative prior distributions is to seek input from an expert. Analysts may conduct interviews with an expert and ask probability-based questions to determine the quantiles of the distribution [9, 10]. This approach has been extensively researched and widely used in practice [11, 12, 13], despite its drawbacks that can lead to inaccuracies in the elicited distributions. Analysts seem to prefer this method, either because they believe it is the only option available or because it can yield positive results for specific applications. However, this is not the only method for eliciting informative prior distributions. There exist alternative techniques that involve experts performing hypothetical decision-making tasks that can address some of the limitations of asking probability-based questions. These techniques involve asking experts specific questions that they can easily comprehend and then having the analyst derive a prior distribution from the expert's responses [14, 15, 16, 17, 18, 19].

The term "Prior Elicitation" typically pertains to the process of obtaining informative prior distributions from experts. However, there are also methods that do not rely on expert judgement to attain informative prior distributions [20, 21], which should be included in discussions alongside methods that leverage expert judgment. Analysts should be aware of all available methods to

acquire informative prior distributions and choose which one to utilise based on their specific task requirements [22]. For this reason, the term elicitation in this text is taken from its root meaning, to obtain something [23]. Prior elicitation methods refer to methods that obtain a prior distribution from a source (be that source human or otherwise).

More information on current elicitation techniques can be found in Chapter 3.

## 2.3 Deep Learning

This thesis explores the process of eliciting expert uncertainty and proposes a new approach that models expert decision-making. To aid experts in their decision-making, they are typically presented with reports or images, which are difficult to model using regular statistical methods. Therefore, deep learning models are suggested as they can handle complex data. Deep learning is a sub-field of Artificial Intelligence (AI) that uses deep neural networks to learn. Understanding the human decision-making process is crucial in the field of AI [24]. These networks are similar to the structure of the human brain and its information processing [25]. The network has layers of neurons that process information and pass it to the next layer until it reaches an output layer. Figure 2.1 illustrates the basic structure of a single hidden layer network with one neuron. In this network, the $X_i$ nodes represent inputs, while their associated parameters (weights, $w_i$, and bias term, $b$) are used at the neuron, $Z$, to multiply and sum over all inputs. The resulting value is then passed through an activation function to produce the output, $\hat{y} = g(Z)$. This basic structure can be built upon to make more complex models with many layers that can process more complex data.

An analyst can use a convolutional neural network (CNN) as a base model for images. A CNN is a neural network that can analyse grid-based inputs,

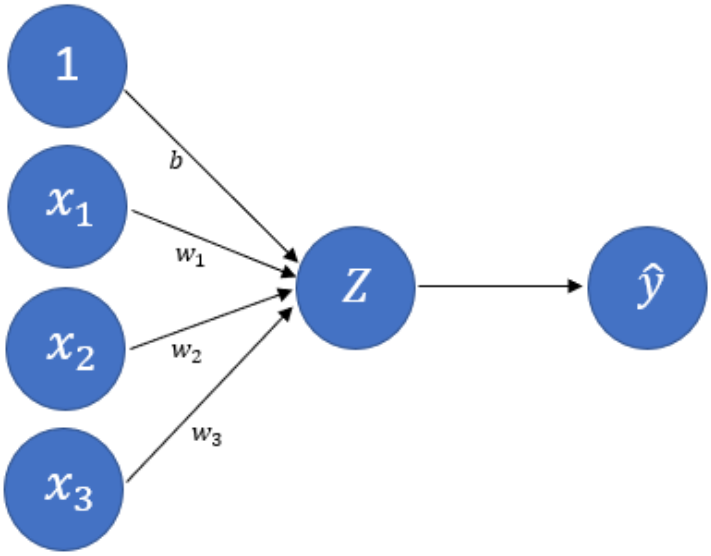



Figure 2.1: Simple Neural Network Architecture

such as image/2D grids of pixels [26, 27]. It consists of convolutional layers, pooling layers, and an output layer. During the convolution step, a filter (a matrix of smaller dimensions than the input) is applied to a section of the input matrix by multiplying element-wise and adding together all the outputs. The process is repeated for other sections of the input matrix, resulting in a smaller dimension matrix representing the input matrix (Figure 2.2). This filter serves as the "weights" of the convolutional layer and is the parameter the model tries to learn. Pooling layers are used between convolutional layers to help stop small changes to the input changing the input of the next layer [27]. Max-Pooling is an example of a pooling layer, which takes the maximum value of a section of the previous layer's matrix. The last layer of a CNN is fully connected (the output layer), meaning that all previous layer nodes are used to calculate the output. This final layer is used to classify the input.

An analyst can use a recurrent neural network (RNN) for modelling decision-making processes that involve assessing text documents. An RNN is a neural network that can analyse sequential data, e.g. sentences and times series data [27]. RNNs have inputted values at each time step and can either output val-

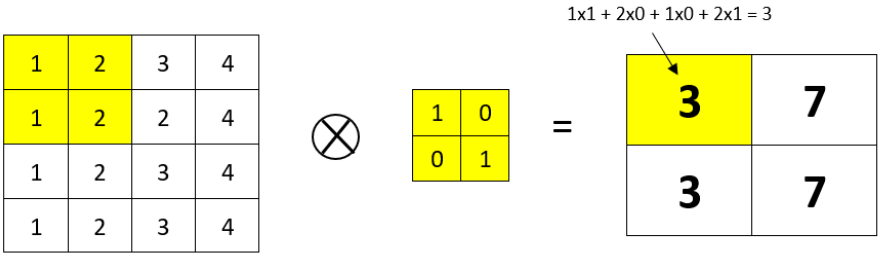



Figure 2.2: Convolution step in a CNN

ues at each time step or have a single output at the end of the sequence. We wish to have only the output at the end of the sequence to elicit a distribution. When each input is processed, past information from previous inputs is passed to the next step (Figure 2.3). An extension of the RNN is the Long Short Term Memory Network (LSTM), which can control what information is being passed through the network. A bidirectional RNN is also helpful, as not only does it have knowledge from the past at a given state but also knowledge from the future; this can be particularly useful for language processing [28].

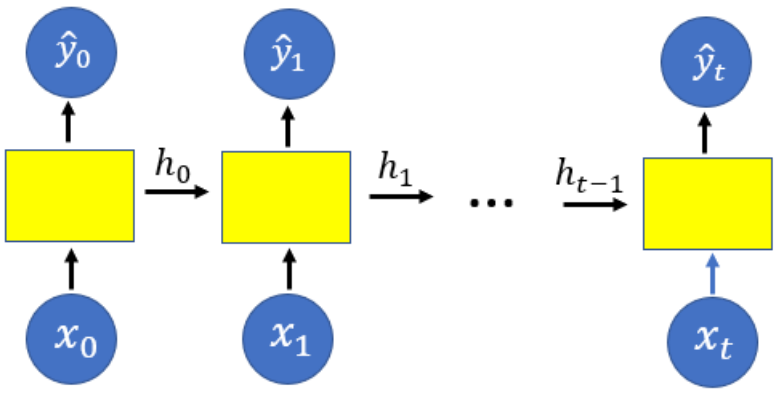



Figure 2.3: Structure of a RNN

Learning/training of the deep learning model is done by what is known as stochastic gradient descent [27]. Back-propagation calculates gradients and then parameters are adjusted in stochastic gradient descent. An analyst will select a loss function and a learning rate for each model. The learning rate is

the factor by which the model will update weights to minimise the loss. The goal is for the model to learn the weights that minimise the loss function while maintaining good accuracy for a given holdout set, the validation data set (so the model is not over-fitting to the training data set). Training data is often split into smaller sets, referred to as batches, of data for the model to process. The entire training set is fed through the network a certain number of times, referred to as epochs. The number of epochs is selected to ensure the model training maintains the balance between minimising loss and not over-fitting.

The field of deep learning is rapidly growing, and it has many different models which can handle many different types of data. This section only introduces the basic structure of two of the main models; many other complex pre-built models may help the analyst produce better results [29, 30]. We suggest the reader read [27] for a more in-depth review of the topic than what we provide in this research. These models make it possible to process complex data. However, we must extend these models to be able to elicit a distribution. These extensions are discussed further in Chapter 5.

### 2.3.1 Probabilistic Deep Learning

This thesis discusses two probabilistic deep learning models, Bayesian neural networks and neural networks that contain Monte Carlo dropout (MC dropout). Bayesian neural network models are neural networks that follow Bayesian principles. First introduced in 1989 [31], Bayesian neural networks require prior distributions to be placed over the weights of the network, resulting in learning a posterior distribution over said weights. This allows a Bayesian neural network to represent uncertainty, which is one reason why they are selected over standard neural networks [32]. The posterior distributions of Bayesian neural networks are often intractable, like other Bayesian models. These posterior distributions can be approximated, similarly to standard Bayesian models, using MCMC methods or variational methods (as dis-

cussed above in Section 2.2). MCMC methods require samples to be taken and update algorithms applied to each parameter of the neural network. Implementing MCMC methods for large neural networks with numerous parameters can be computationally expensive and often infeasible [33]. Variational Bayes (also known as variational inference) is the preferred method for approximating the posteriors in Bayesian neural networks because it is computationally easier to compute approximations; it requires less time and memory than MCMC methods [33]. Standard back-propagation algorithms are not suitable for the use of variational inference in Bayesian neural networks due to the stochastic nature of the weights. Alternative methods have been put in place to perform back-propagation [34, 35].

MC dropout [36] is another probabilistic deep learning feature that allows an analyst to output a distribution that captures uncertainty. Dropout is a component of a network that randomly "drops out" nodes from a layer during calculations. An analyst will select the value of the dropout probability, $q_i$, when building the model. For each pass through the network, every node has a probability $q_i$ of being excluded from calculations. A neural network with MC dropout can produce distributions that capture uncertainty by incorporating dropout into each layer of the network and then running inputs through the network multiple times, with each run through the network dropping out different nodes. MC dropout is considered a Bayesian approximation [36] of a neural network and is less computationally expensive to implement than standard Bayesian neural networks.

This thesis aims to quantify expert uncertainty by modelling decision-making data. Bayesian neural networks and MC dropout are both suitable for capturing uncertainty and are discussed further in Chapter 5.

## 2.4 Uncertainty

The goal of this research is to elicit expert uncertainty on the probability of some event, $E$. There are two types of uncertainty that we discuss in this thesis; aleatoric uncertainty and epistemic uncertainty [37]. Aleatoric uncertainty is the uncertainty caused by natural random variation and is irreducible (or cannot be reduced in the near future [37]), such as a coin flip. Each time we toss a "fair" coin, we are uncertain of its outcome. Epistemic uncertainty is caused by a lack of knowledge and is reducible; the more knowledge or information we obtain, the more certain we become. In this thesis, we classify the type of uncertainty based on the event, $E$, itself. If the goal is to elicit the aleatoric uncertainty of $E$ from an expert, then we expect the uncertainty of $E$ to be irreducible at the time of elicitation and to be caused by randomness. If the goal is to elicit epistemic uncertainty, then we expect the uncertainty of $E$ to be reducible and caused by some lack of skill or information.

It is important to note that in other research fields, these concepts have different terms assigned to them. Occasionally, aleatoric uncertainty and epistemic uncertainty are known as objective uncertainty and subjective uncertainty, respectively [38]. Although they have the same meaning as above, we do not want to confuse the reader by using these terms, as the word "subjective" means anything that is based on personal opinion, so any uncertainty that we elicit from experts would be considered subjective. In deep learning, aleatoric uncertainty and epistemic uncertainty are referred to as data uncertainty and model uncertainty, respectively [39], as at the time of model building, the uncertainty in the data cannot be reduced, yet the model uncertainty can be reduced by improving the model. However, in this research, the uncertainty we wish to elicit is in the decision-making data that we put into our deep learning model, so our uncertainty in the data can be either aleatoric or epistemic, dependent on the application.

It is important to keep this distinction in mind when reading through the

remainder of this thesis.

# References

[1] M Granger Morgan. Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National academy of Sciences*, 111(20):7176–7184, 2014.

[2] Bryan D Jones. Bounded rationality. *Annual review of political science*, 2(1):297–321, 1999.

[3] Robert JB Hutton and Gary Klein. Expert decision making. *Systems Engineering: The Journal of The International Council on Systems Engineering*, 2(1):32–45, 1999.

[4] James Shanteau. Psychological characteristics and strategies of expert decision makers. *Acta psychologica*, 68(1-3):203–215, 1988.

[5] Isabella Fornacon-Wood, Hitesh Mistry, Corinne Johnson-Hart, Corinne Faivre-Finn, James PB O'Connor, and Gareth J Price. Understanding the differences between Bayesian and frequentist statistics. *International journal of radiation oncology, biology, physics*, 112(5):1076–1082, 2022.

[6] Eric-Jan Wagenmakers, Michael Lee, Tom Lodewyckx, and Geoffrey J Iverson. Bayesian versus frequentist inference. *Bayesian evaluation of informative hypotheses*, pages 181–207, 2008.

[7] M Jésus Bayarri and James O Berger. The interplay of Bayesian and frequentist analysis. 2004.

[8] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[9] Lionel A Galway. Subjective probability distribution elicitation in cost risk analysis: A review. 2007.

[10] David Jenkinson. The elicitation of probabilities: A review of the statistical literature. 2005.

[11] Anthony O'Hagan, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. Uncertain judgements: eliciting experts' probabilities. 2006.

[12] Anthony O'Hagan. Expert knowledge elicitation: subjective but scientific. *The American Statistician*, 73(sup1):69–81, 2019.

[13] Sindhu R Johnson, George A Tomlinson, Gillian A Hawker, John T Granton, and Brian M Feldman. Methods to elicit beliefs for Bayesian priors: a systematic review. *Journal of clinical epidemiology*, 63(4):355–369, 2010.

[14] Robert L Winkler. The quantification of judgment: Some methodological suggestions. *Journal of the American Statistical Association*, 62(320):1105–1120, 1967.

[15] Carl S Spetzler and CS Stael Von Holstein. *Probability encoding in decision analysis*. Stanford Res. Inst., 1972.

[16] Ali E Abbas, David V Budescu, Hsiu-Ting Yu, and Ryan Haggerty. A comparison of two probability encoding methods: Fixed probability vs. fixed variable values. *Decision Analysis*, 5(4):190–202, 2008.

[17] Enrico Cagno, Franco Caron, Mauro Mancini, and Fabrizio Ruggeri. Using ahp in determining the prior distributions on gas pipeline failures in a robust Bayesian approach. *Reliability Engineering & System Safety*, 67(3):275–284, 2000.

[18] Robert T Eckenrode. Weighting multiple criteria. *Management science*, 12(3):180–192, 1965.

[19] Ward Edwards and F Hutton Barron. Smarts and smarter: Improved

simple methods for multiattribute utility measurement. *Organizational behavior and human decision processes*, 60(3):306–325, 1994.

[20] S James Press. *Subjective and objective Bayesian statistics: Principles, models, and applications*, volume 590. John Wiley & Sons, 2009.

[21] Joseph G Ibrahim, Ming-Hui Chen, et al. Power prior distributions for regression models. *Statistical Science*, 15(1):46–60, 2000.

[22] Joseph Kadane and Lara J Wolfson. Experiences in elicitation: [read before the royal statistical society at a meeting on'elicitation 'on wednesday, april 16th, 1997, the president, professor afm smith in the chair]. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):3–19, 1998.

[23] Cambridge University Press. *Cambridge Academic Content Dictionary*. Cambridge Academic Content Dictionary. Cambridge University Press, 2008.

[24] Jean-Charles Pomerol. Artificial intelligence and human decision making. *European Journal of Operational Research*, 99(1):3–25, 1997.

[25] Kevin Gurney. *An introduction to neural networks*. CRC press, 2018.

[26] Yann LeCun et al. Generalization and network design strategies. *Connectionism in perspective*, 19:143–155, 1989.

[27] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[28] Charu C Aggarwal et al. *Neural networks and deep learning*. Springer, 2018.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[30] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[31] Naftali Tishby, Esther Levin, and Sara A Solla. Consistent inference of probabilities in layered networks: Predictions and generalization. In *International Joint Conference on Neural Networks*, volume 2, pages 403–409, 1989.

[32] Hao Wang and Dit-Yan Yeung. A survey on Bayesian deep learning. *ACM Computing Surveys (CSUR)*, 53(5):1–37, 2020.

[33] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[34] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015.

[35] José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *International Conference on Machine Learning*, pages 1861–1869. PMLR, 2015.

[36] 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.

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

[38] Fabio Campos, Andre Neves, and Fernando M Campello de Souza. Decision making under subjective uncertainty. In *2007 IEEE Symposium*

*on Computational Intelligence in Multi-Criteria Decision-Making*, pages 85–90. IEEE, 2007.

[39] 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.

# Chapter 3

# Methods for Eliciting Informative Prior Distributions: A Critical Review

Eliciting informative prior distributions for Bayesian inference can often be complex and challenging. While popular methods rely on asking experts probability-based questions to quantify uncertainty, these methods are not without their drawbacks, and many alternative elicitation methods exist. This paper explores methods for eliciting informative priors categorized by type and briefly discusses their strengths and limitations. Most of the review literature in this field focuses on a particular type of elicitation approach. The primary aim of this work, however, is to provide a more complete yet macro view of the state of the art by highlighting new (and old) approaches in one clear, easy-to-read article. Two representative applications are used throughout to explore the suitability, or lack thereof, of the existing methods, one of which highlights a challenge that has not been addressed in the literature yet. We identify some of the gaps in the present work and discuss directions for future research.

## 3.1 Introduction

Bayesian inference can be thought of as the process of updating prior knowledge once the data, $y$, has been observed. Bayes' rule gives

$$p(\theta|y) \propto p(\theta)p(y|\theta), \tag{3.1}$$

where $p(\theta)$ is the prior distribution on parameters of interest, $\theta$, and $p(y|\theta)$ is the likelihood function for data, $y$. The prior distribution could be considered to be *informative*, that is, elicited based on available information and beliefs, or *non-informative*, where no such prior information or beliefs may be available. If there is a large amount of data, then the inference will be influenced more by the likelihood function. The opposite of this is also true; if there is a limited amount of data, information from the likelihood will be weak, and the inference will be influenced more by the prior distribution. In this case, the choice of the prior, $p(\theta)$ is of greater importance. Ideally, an informative prior would be used because the decision-making and inference for this problem will primarily rely on the prior.

A simple example of when an informative prior may be required is as follows,

**Example 1.** *A pharmaceutical company wishes to predict whether or not patients will develop blood clots after taking a new drug they are developing. Let* $Y \in \{1, 0\}$*, therefore* $Y \sim Bernoulli(\theta)$*, where* $\theta$ *is the probability of a patient developing blood clots after taking the new drug. Data on existing patients is limited due to the new formula of the molecule. The goal is to obtain a prior distribution on* $\theta$*, which can be used for further research.*

For Example 1, a common approach for obtaining an informative prior distribution would be to elicit information from an expert, i.e., to perform expert *prior elicitation*. The process may require multiple individuals with different

expertise and roles. A set of standard definitions of these, as used throughout this paper, can be found in Table 4.1. Although obtaining information from experts may seem straightforward at first glance, it often is anything but. Several challenges may arise, complicating the process. We discuss the key challenges below.

Table 3.1: Definitions

| Name | Description |
|---|---|
| *Prior Elicitation* | This paper refers to prior elicitation as the process of obtaining knowledge from a source to form a prior distribution which can be used for further Bayesian analysis. Also referred to in texts as "Probability Encoding". |
| *Expert* | An expert is an individual who has extensive knowledge on a certain subject matter. Also referred to some in texts as a "judge". |
| *Analyst* | An analyst is an individual who performs the task of forming a prior distribution using prior elicitation techniques. |
| *Facilitator* | A facilitator is an individual who performs the task of eliciting knowledge. In some cases, the Facilitator and the Analyst may be the same individual. |

### 3.1.1 Challenges in the Elicitation Process

#### Cognitive Biases

[1] pointed out that natural thought processes may create inaccuracies in the obtained priors by introducing cognitive biases not observable by the analyst. As these cognitive biases are present in the mind of the expert, it is often complicated to adjust for them in the elicited distribution. While many methods have been proposed to mitigate these biases, it could be argued that the biases can never be completely eliminated. An overview of key cognitive bi-

ases and bias reduction methods for prior elicitation are outlined in Section 3.2.

### Using Multiple Experts to Elicit Priors

It is good practice to use multiple experts to elicit a prior distribution [2, 3]. Doing so may ensure that the elicited distribution is better aligned with the entire field of interest and that individual biases present in one expert are reduced. The process of obtaining a prior distribution that aggregates the inputs from multiple experts provides a new set of challenges. Two main approaches to aggregating the prior beliefs of multiple experts have been proposed. Each approach has its benefits and drawbacks; these are discussed in Section 3.6

### Dealing with Little or No Updating Data

Many applications exhibit non-trivial complications. These include applications where there is unlikely to be much data available to update prior beliefs as needed when using Bayes' theorem. The absence of data for updating can occur because the data collection mechanism is too complex or expensive or because the data pertains to an occurrence of a highly undesirable event (e.g., a major accident or a security threat, etc.). In such situations, the decision-making will be predominantly based on the elicited prior distributions since the data will have very little or no influence. A simple example is a problem in the security field:

**Example 2.** *A company wishes to predict whether or not its new security system will fail,* $Y \in \{1, 0\}$*. Therefore* $Y \sim Bernoulli(\theta)$ *where* $\theta$ *is the probability of a system fail. For inference, data on how well the system performs is not present because that would mean security threats have already happened. If the security threat was to materialize, this will result in a single observation on* $Y$*. While* $p(\theta|y)$ *can now be computed, it is likely to be only slightly different to* $p(\theta)$*. Further, in practice, such a breach of the security system will likely result in changes being made so as to prevent such occurrence in the future.*

*This may mean that the circumstances on which the original prior was elicited are no longer present and therefore eliciting a fresh set of prior distributions could be considered more appropriate rather than using the posterior obtained on the original prior.*

Another important facet of such applications is that while it may not be possible to update the prior beliefs on $\theta$, there likely exists some background/ additional information, $x$, that may help in eliciting $p(\theta)$. Hence, $p(\theta)$ can be written as a function of $x$ and other model parameters $\beta$. That is, $p(\theta) = g(x, \beta)$ (see Figure 3.1). This could be thought of as a standard regression problem; however, there may be situations where background information, $x$, is disjointed and can contain different data types; such as image data, text data and expert opinion. Also, some of this information may only be vaguely relevant. Therefore, treating this example as an elicitation process may seem more reasonable. It is important to note that this type of application does not seem to be discussed much in the prior elicitation literature. Yet, it represents a practical decision theoretic problem whose solution may rely on the elicited probabilities.

**Example 2 (contd.).** *The company brings in experts to provide information on system failure. The company provides the experts with all information on their system (e.g., system type, location,...). The experts are able to give opinion on system failure based on the information provided and historical information they have obtained from other establishments and different security systems, $x$. Because each establishment is unique, what happened elsewhere may only be vaguely relevant to this company. The experts would aim to utilise all this related information to elicit a prior distribution on system failure.*

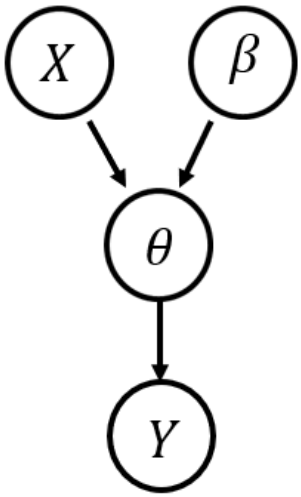



Figure 3.1: Graph showing the Model Structure.

### 3.1.2 Motivation

This paper outlines methods used for eliciting informative priors and the benefits and drawbacks of these methods. Review articles in this field tend to focus on direct interrogation methods (Section 3.3.1) (see for example [4, 5, 6] and [7]), without exploring the other available methods presented in the literature. Other review articles focus on "good practice" and answering questions that may arise in the overall elicitation process but do not expand on actual methods for eliciting priors [8].To the best of our knowledge, there has been no work to date that outlines all of the available approaches and provides an overview of the current state of art. In this paper, we aim to provide such an overview and also to identify some of the gaps in the current literature.

### 3.1.3 Outline

The elicitation methods are classified into three approaches: Interrogation Methods (Section 3.3), Graphical/Visual Methods (Section 3.4), and Historical Information Methods (Section 3.5). Interrogation methods are elicitation methods that involve asking an expert questions to obtain a probability distri-

bution. Graphical/Visual Methods involve plots of distributions or data that the expert can visualise and thus compare with their individual knowledge. Methods that are labelled Historic Information Methods are those that may not rely on expert input and instead utilise information from past research studies. An overview of cognitive biases is presented in Section 3.2. Techniques for combining information from multiple experts are reviewed in Section 3.6. The paper concludes in Section 3.7, by discussing some of the persistent challenges of informative prior elicitation and proposing further research paths.

## 3.2 Cognitive Biases in Expert Knowledge Elicitation

The following are considered key cognitive biases that may be present in the expert elicitation process (for a comprehensive list of issues that arise in human judgement, see [9]).

- *Judgement by Representativeness*; this may be present with questions such as "What is the probability that an object A belongs to a class B?" [10]. Often with these questions, the expert will focus on the conditional probability and completely ignore the unconditional probabilities. [1] outline a clear example of Judgement of Representativeness in assigning the probability of an individual having a certain job, when the expert knows the individual's personality. The expert ignores the number of people in the population in the jobs and instead relies solely on the conditional probability, that is, they judge the probability of an individual having a certain job given the fit between the individual's personality and the job [1].

- *Judgement by Availability*; this is when an event is given a higher probability based solely on the fact that it occurred more recently for the expert [1], so they are able to easily bring it to mind. Considering Example 1, an expert may have recently come across a case where a patient developed blood clots from the new drug. Even though the example may be rare, because it is fresh in their mind, the expert may believe that the risk of developing blood clots is higher, in turn, influencing the elicited prior.

- *Anchoring and Adjustment*; here, an expert is given some value (the anchor) and adjusts it to achieve what they think is the correct value. Experiments have shown that an expert who starts with a higher anchor is more likely to give higher estimates than experts who start with lower anchors [1]. In Example 1 if the facilitator states a value for the mean of the probability of a patient obtaining blood clots, the expert may then adjust this value to what they think is suitable, producing results that are inaccurate due to the initial anchor value presented to them.

- *Over Confidence* of the expert; Overconfidence is easily observed in studies where experts were asked for an interval [11]. For example, where experts were asked for 95% probability intervals, it was found that as little as 65% of those intervals actually contained the true values [2]. Thus, the expert often narrows down their judgment of uncertainty, displaying overconfidence. [12] gives an overview and comparison of methods that help evaluate an expert's overconfidence, which may be helpful to researchers.

- *Range Frequency*; this bias is seen when an expert is asked to assign the probability of each category and they evenly assign probabilities between

categories [13, 2]. This means that categories believed to be more likely are given less probability than required and others are given more, creating a significant bias. For Example 1, an example could be when asking for the probabilities for a male developing blood clots and a female developing blood clots; the expert assigns equal probability to each group.

- *Expert Fatigue.* When an expert is required to take in and store copious amounts of information or complex information to complete the elicitation process, the process can become cognitively tiresome and time consuming for the expert [14]. Analysts can reduce this burden by simplifying the model's parameters, by splitting the task up into different parts and by explaining complex quantities in layman's terms for the expert [14]. All the while, analysts must keep in mind the trade-off between easing expert fatigue and still having models that represent the real-world complexities. Expert fatigue varies from expert to expert and is heavily dependent on the level of expertise needed for the task [15].

Because cognitive biases are present in any human thought process, one goal of the expert elicitation methods should be to reduce these biases as much as possible. These cognitive biases can be reduced by changing the types of questions asked [16]. However, the way questions are asked may still affect the elicited prior. Another approach to reducing cognitive biases is a calibration technique, first introduced by [17]. The goal of this technique is to calibrate the expert's opinion based on the biases that they have shown in past opinions they have provided [18, 19]. The technique involves using information from past estimates given by the expert where the true value is now known. [19] use hierarchical Gaussian processes to model the expert's biases from their past estimates. The models are trained on the calibration data (historical estimates and the true values) and then used to correct for bias in the new estimates

given by the expert [19]. Although a promising technique, the obvious disadvantage is, when calibration data is limited or non-existent, that this technique becomes impossible.

Others have suggested that the best way to address cognitive biases is to give the expert feedback and make them accountable for their responses [20]. Training them to give more appropriate responses in the future [18]. Cognitive biases have been outlined in detail in terms of expert prior elicitation in [10, 2, 8].

Methods to elicit an informative prior, discussed in Sections 3.3,3.4 and 3.5, aim to reduce such biases; however, the biases above may still occur in each method.

## 3.3 Interrogation Methods

Interrogation methods are those in which an expert is interviewed to obtain knowledge on a parameter of interest [21].

### 3.3.1 Direct Interrogation

Distribution or probability based questioning is the most common form of prior elicitation [5, 7]. Classified as direct interrogation methods, these techniques use a questionnaire or interview with an expert, with questions that are specifically related to probabilities and/or distributions. [22] provide a systematic review of papers that involve direct interrogation methods for prior elicitation, outlining specific papers and the questions used. [16] outline procedures for researchers to create their own questions for their specific task and refine them where necessary, taking into account potential biases that may arise. Techniques for direct interrogation include:

- **Questions on Probabilities**: Asking questions on probabilities that could directly link to the Cumulative Distribution Function (CDF) or Probability Density Function (PDF) [6]. For instance, when eliciting a prior for Example 1, a possible question could be "What is the probability that the proportion of those who get blood clots is less than or equal to 0.3?", i.e., $P(Y \leq 0.3)$. Once a few points have been collected, the researcher can obtain an outline of the CDF or PDF.

- **Questions on Distribution Quantiles**: Asking questions on the quantiles of the CDF or PDF [23]. In the case of Example 1, a possible question could be "At what value is the probability of a patient obtaining blood clots equally likely to be less than or greater than that value?" ( i.e., estimate the median).

Under each of theses techniques, the facilitator can either ask the expert to give a value for the variable given a fixed probability or ask the expert to give a probability for a fixed value or ask questions where to expert is expected to give both probability and value [24]. Some analysts prefer to ask questions directly on the parameter of interest. This is known as *Structural Elicitation* [25], this should not be confused with *Structured Elicitation* which is a step-by-step process of elicitation which may contain *Structural Elicitation*. [6] (Chapter 5) give an extensive review of such techniques, detailing acceptable methods and how to use the expert's responses to form a probability distribution (also summarised in [10]). Tools such as SHELF [26] and MATCH [27] are suitable for forming distributions from this line of questioning. These tools also provide instant visual feedback of the expert's opinions in the forms of fitting a distribution which has been found to be beneficial to the elicitation process [6, 22]. Techniques which directly involve graphical and visual components are discussed separately in Section 3.4 of this paper.

[28] argued that in some cases parameters are arduous to think about contextually, which can create difficulties in the elicitation process; as the expert will find it difficult to form a belief on the parameter's behaviour. Therefore, questions should be asked on observable values, i.e., the predictive distribution. This approach is known as *Predictive Elicitation*. While some experts may find it hard to understand the real-world applications of model parameters, experts should, by definition, have copious knowledge on the field of expertise data. Predictive Elicitation has been explored further in [25, 29] and [30], with [25] providing a comparison of Structural and Predictive Elicitation, showing examples of when each technique is appropriate. [25] emphasise that creating a predictive elicitation process is task specific and must be adapted for each task undertaken by the analyst. Predictive elicitation techniques are also more complex and time consuming to perform than structural techniques [25]. It is worth noting that these techniques essentially treat prior elicitation as an inference problem. Consider a basic example for a regression model, a facilitator would ask the expert about their uncertainty around the dependent variable given different values of the response variable. Once the uncertainty is captured, the analyst would then have to infer the range of values for the model parameters and their prior distributions that would be consistent with the expert's elicitation [25].

Applying direct interrogation methods to Examples 1 and 2 may be appropriate if the expert has good comprehension of statistical concepts needed. But even if they do, common cognitive biases, such as *Expert Fatigue*, *Judgement by Representativeness*, *Judgement by Availability*, *Anchoring and Adjustment*, *Over Confidence* and *Range Frequency*, may still affect the process. This is when calibration techniques, as discussed in Section 3.2 may be of use.
Direct Interrogation Methods also have the additional obstacle of first requiring the facilitator to teach the expert to understand statistical concepts, which can be complex. Even once the elicitation process has started, the expert may

still not understand key concepts needed, nor be able to apply them to form an accurate probability statement. [25] state *"The goal of elicitation, as we see it, is to make it as easy as possible for subject-matter experts to tell us what they believe, in probabilistic terms, while reducing how much they need to know about probability theory to do so"*. Hence, other types of interrogation methods exist that can reduce the statistical knowledge required to complete the task of eliciting a prior from experts, discussed next in Section 3.3.2. Some other relevant methods are discussed in Sections 3.4 and 3.5.

### 3.3.2 Indirect Interrogation

Probability-based questioning is not the only form of questioning in interrogation techniques. There are approaches to form distributions from other types of questioning, known as indirect interrogation techniques.

- **Betting method:** In this method, a series of bets are placed to form a distribution [31]. The process starts by presenting the expert with two bets and asking the expert to select the bet based on the event they believe is more likely to occur. An example from [31] is as follows: Let **Bet One** be "*Win $A if Event E occurs, lose $B if E doesn't occur*" and **Bet Two** be "*Win $B if E doesn't occur, lose $A if E occurs*". The expected values of each will be **Bet One**: $Ap - B(1-p)$, **Bet Two**: $B(1-p) - Ap$. If the expert chooses Bet One , then for this expert $Ap - B(1-p) \geq B(1-p) - Ap \Rightarrow p \geq B/(A+B)$. By letting Event, E, be any combination on the real line, analysts can elicit information on a probability distribution. Another approach is to use a probability wheel, split into two colours, where an expert can select between two scenarios; either they win $\$X$ if the probability wheel lands in a grey area or they win $\$X$ if an event occurs [24]. The facilitator then changes the area of the grey area and/or the event specifications to present a new question

to the expert. This is repeated until the expert finds both situations equally likely [24, 32]. The number of bets to form an appropriate distribution could be relatively large, causing expert fatigue; to avoid this, betting can also be used alongside direct interrogation methods to check the certainty of assessed probabilities [31].

Using the betting method for both Example 1 and 2 may require extensive thought and time. Setting up suitable bets will be the first obstacle, but then the analysts must create enough bets to form a more complex distribution, which can be time consuming. Betting situations may prove more useful in the case where an analyst needs to elicit an individual probability rather than a complete distribution.

- **Pair-wise Comparison:** In this approach, experts compare pairs of categories and quantify their beliefs regarding which category is more likely. This action is carried out for each possible pair in the study. An Analytic Hierarchy Process (AHP) [33, 34] is a pair-wise comparison process that can also be used for prior elicitation [35]. For comparisons, experts are asked to give a number on a scale (from 1 - 9) based on which event they believe is more likely to occur. Using Example 2, an AHP approach could be used to obtain prior distributions for different locations of sites where the security system is used. For example, let building location be the set {"Country", "City Centre", "Suburbia", "Industrial"}. The expert can then compare which system is more likely to fail for each pair of building locations. This forms an opinion matrix, as shown in Table 3.2.

  Line four in the opinion matrix reads [35] that the expert believes buildings in Industrial areas are very strongly more probable to have security failure than buildings in Country areas and strongly more probable to have security failure than buildings in the City Centre and Suburbia ar-

Table 3.2: Opinion Matrix for the AHP Method for Building Location in Example 2

| Building Location | Country | City Centre | Suburbia | Industrial |
|---|---|---|---|---|
| Country | 1 | 1/3 | 1/3 | 1/7 |
| City Centre | 3 | 1 | 1 | 1/5 |
| Suburbia | 3 | 1 | 1 | 1/5 |
| Industrial | 7 | 5 | 5 | 1 |

eas. By taking the maximum eigenvalue of the opinion matrix above and calculating the associated eigenvector, the analyst can obtain the vector of weights of the building areas [33, 34], where the weights are the expert's security failure likelihood (or propensity-to-failure in [35]) for each building location. [35] then calculated the mean and standard deviation of the expert's opinions and used these to obtain the parameters of a Gamma distribution for each class.

Although there may be some use for this method in the case of Example 2 (where there is potential for sub-classes and experts have observed examples from each class), for Example 1, reducing the case down to specific sub-classes may not necessarily be possible. This method obviously becomes impractical when there are no sub-classes in the research task.

- **Ranking/Rating Method:** in this method, experts are asked to rank or give a rating on the likelihood of events presented to them [36, 37]. [38] give an appropriate example of this method in terms of decision analysis. In the application of understanding terrorist attacks, they asked intelligence officers to rank the attractiveness of selected potential targets. They then used probabilistic inversion and Bayesian density estimation to elicit the distributions [38]. This method can also be used for prior elicitation under time pressure [39]. A ranking method may be hard to implement for Example 1, where the analyst may find it difficult to split the problem into sub-groups that would be rank-able. For Example 2,

this method may work quite well. The analyst has the additional information on the property so they could come up with other properties that have different attributes to the property of interest and ask the expert to rank a list of properties.

These techniques are designed to avoid having to teach the experts about statistical concepts, which can be a drawback of direct interrogation methods [38]. They can also simplify the process; however, analysts need a strong grasp on the subject matter to be able to form a suitable set up for elicitation (e.g., forming sub-groups and different attributes which may effect the elicited distribution). Indirect interrogation methods also clearly remove the cognitive biases of *Judgement of Representativeness* and *Range Frequency*, by not asking questions relating to probabilities. However, other cognitive biases may still be present.

## 3.4 Graphical/Visual Methods:

Graphical/Visual Methods are methods that involve graphical representations of the probability distribution and/or data that may be used to form a distribution.

- **Trial Roulette method**: First established by [40], this method involves a graphical representation of the parameter space of the parameter of interest being split into subsections. The expert then assigns blocks to the subsections, essentially building their own probability distribution. The trial roulette method has been used in practice in [41].

  In Example 1, the online tool MATCH [27] can be used to elicit the expert's belief via the Trial Roulette Method. The MATCH tool for the Trial Roulette method is shown in Figure 3.2. In the particular example

scenario considered in Figure 3.2, the expert has placed the majority of the blocks towards the zero end of the scale, meaning they believe the probability of someone developing blood clots is small. They still, however, have some uncertainty, having placed blocks as far out as 0.5 to cover this uncertainty regarding the location of the true value. In this scenario, the tool has found the $Gamma(1.03, 6.25)$ distribution to be an appropriate fit for the expert's distribution. The user may change this distribution if desired and select (say) a Scaled-Beta which may be considered to be more appropriate given its support and conjugacy to the Bernoulli likelihood.

Although the Trial Roulette method is a great visual tool, it still requires the expert to have sufficient statistical knowledge to be able to place the blocks to form an appropriate distribution. Also, the user needs to understand probability distributions to avoid inaccuracies in the default selections. For instance, in the example discussed above, conceptually, a Gamma distribution is not appropriate for proportions as it can take values greater than one.

- **Graphical Prior Elicitation**: [42] outline a simple graphical tool for prior elicitation in univariate models. This tool initially teaches the expert about the different distribution types and what effect the distribution parameters have on them. Explaining this method with Example 1, the user begins by selecting an initial model distribution; the distribution on $Y$, a Bernoulli distribution. The tool automatically selects a conjugate prior distribution on the parameter, a Beta Distribution. The user will then select the expected number of successes out of 100, automatically generating plots of what the number of successes vs. failures would look like, see Figure 3.3. Changing the number of successes, the expert can then see how the resulting plots will change. This can be thought of as a Predictive Elicitation technique, as the user is observ-

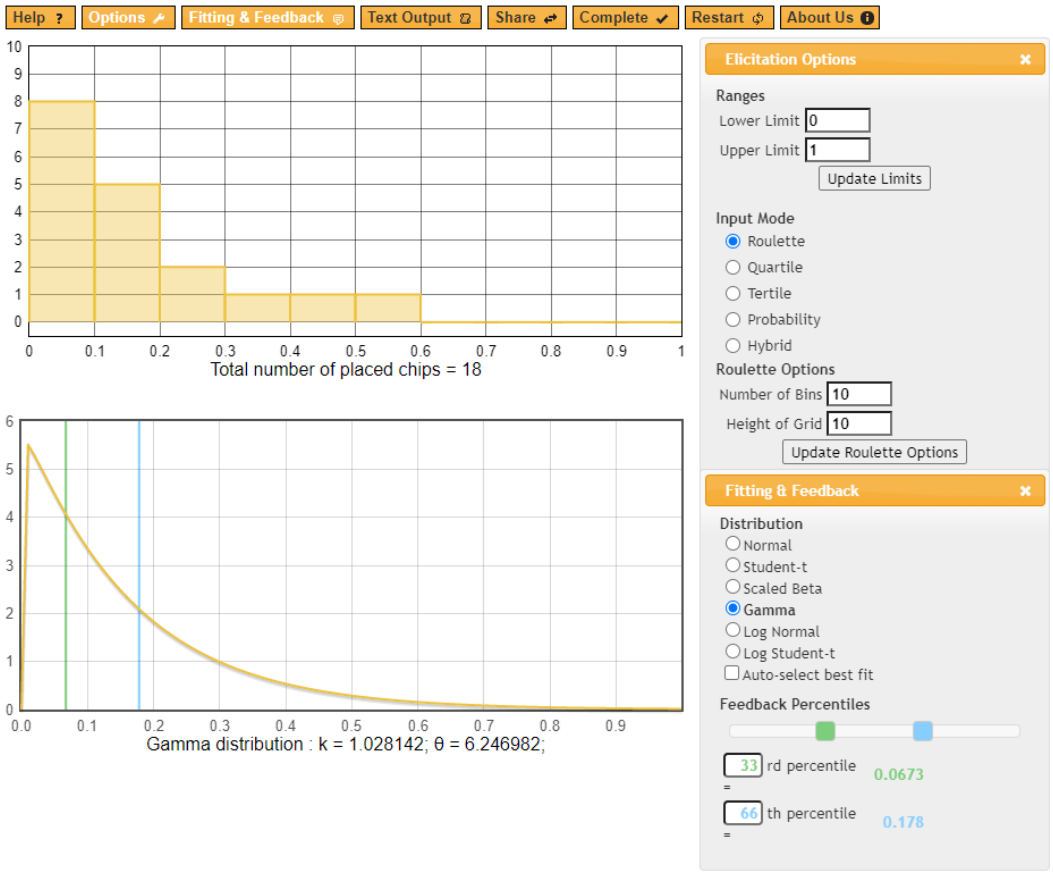



Figure 3.2: Online Tool MATCH displaying the Trial Roulette Method

ing plots of the data, not the parameter. After training the expert, the tool then displays a collection of sample data histograms, of which the expert must select those that are similar to what they believe the data would be. From there, the prior distribution is formed from the selected histograms. [42] suggest this method simplifies the statistical knowledge required to produce a prior distribution, significantly reduces the time of training the expert and the processing time of forming a prior distribution. However, experts still need partial statistical understanding to complete the task of outputting a prior that is consistent with their beliefs.

- **Interactive Excel Spreadsheet**: Starting with a plot of a selected prior distribution (for example, the user may have selected a Gaussian distribution), this method involves the expert moving the "sliders", in


Figure 3.3: Training Step of the Graphical Elicitation Method.

the corresponding excel spreadsheet, to see the effect of changing the value of the parameters on the prior distribution [43]. This approach allows the expert to adjust the parameters until they find their ideal prior distribution. For the case where parameters are not "intuitively meaningful" [43], the interactive spreadsheet, used in this method can show multiple plots of the observable quantity's distributions, allowing the user to see the effects of changes to these distributions when parameters of the original prior distribution are changed. This variation can be thought of as a Predictive Elicitation form of this method.

This spreadsheet based method is similar to the Graphical Prior Elicitation Method in that it allows for experts to first learn the effects of distribution parameters on the given distribution. Also, like the Graphical Prior Elicitation Method, experts still need to have comprehensive knowledge of statistical concepts to be able to understand the distributions themselves, including knowledge of which distribution to initially select. [43] provide example spreadsheets in their supplementary materials and give an overview of different applications where this technique

has been applied.

- **Computer Simulated Data**: [44] introduce a method which involves a plot of simulated data. They illustrate two techniques; in the first technique, the expert observes a plot of simulated data and responds as to whether or not this data could be from a real data set. The second is where pairs of simulated data are compared to one another. Again, this can be thought of as a Predictive Elicitation technique. From there, the expert's responses are used to train a Gaussian Process (GP) classifier that captures the expert's belief [44]. Exploring the first technique further, an expert is shown a simulation $Y_\theta \sim p(Y|\theta, M)$, conditional on model parameters $\theta$, drawn from model $M$. The response from the expert is binary, $z_E = 1$ or 0. These responses are then used to train a GP classifier, $C$, to model $p(z|\theta)$ [44]. Using Bayes theorem, Thomas et al. describe how they can estimate $p(\theta|z = 1, C, z_E, M)$. GP classifiers work by placing a GP prior over a latent function, $f$, and then applying a logistic function, $\sigma$, to $f$ to obtain a prior for class probabilities, $\pi$, that is $\pi(x) = \sigma(f(x))$.

  This method is similar to Graphical Elicitation. The main difference is the selection of the prior distribution type, as this method does not assume a conjugate prior. It is noteworthy that in this method, the expert does not need a strong knowledge of statistical concepts to obtain an accurate prior distribution, as they should have a strong understanding of subject matter data.

Although graphical methods are helpful in providing instant visual feedback to the expert, and, in some cases, simplifying the statistical knowledge required to perform the task, they are still subject to the cognitive biases discussed in Section 3.2. The Computer Simulated Data method and the Graphical Elicitation Method, are interesting approaches, because they aim to

generate a distribution by modelling the decisions of the expert. This may be an interesting topic for further research.

## 3.5 Using Historical Information

Methods discussed thus far aim to elicit the uncertainty that the expert has into a probability distribution. Thus, they could be considered to be similar to a standard translation tool, with the objective of translating the expert's knowledge, including their quantified uncertainty, into a probability distribution. In practice, a standard translation tool that translates a text from one language into another, requires the text to be available in the first place. In the same way, the prior elicitation methods discussed so far require the expert to quantify their uncertainty first. However, there are applications where relevant historical data exists, and the analyst may prefer to quantify uncertainty using such data instead. In this section, methods that elicit prior distributions using available past research are reviewed. The main theme of these methods is that they require minimal human intervention to obtain a prior distribution.

- **Utilising Historical Research Posteriors:** One way to obtain a prior without expert intervention is from similar historical research that provides a posterior distribution. This posterior distribution can then be used as a prior distribution in a new study [45]. However, many similar studies will not provide a concrete posterior distribution, but may have similar data.

- **Power Prior:** The power prior [46] provides a way to obtain a prior from historical data, $D_0 = (n_0, y_0, X_0)$, where $n_0$ denotes the sample size, $y_0$ denotes the response vector and $X_0$ denotes the matrix of the covariates. In a basic form, the power prior is the likelihood of the

historical data, $L(D_0|\theta)$, to the power of a scalar, $a_0$, multiplied by an initial prior $\pi_0(\theta|c_0)$,

$$\pi(\theta|D_0, a_0) \propto L(D_0|\theta)^{a_0}\pi_0(\theta|c_0). \tag{3.2}$$

The scalar, $a_0$, quantifies the uncertainty in the data, $D_0$. It allows the analyst to control how much the historical data influences the prior. The initial prior parameter, $c_0$, allows the analyst to control the influence of the initial prior on the prior, i.e, how much information from the initial prior is influencing the power prior. As selection of these parameters can change the final power prior, analysts should take careful consideration when deciding these values [46]. Other variations of the power prior [47] may be of use to some practitioners, along with methods to form priors from not only historical data but also historical research that only provides summary statistics of the data [48].

These methods help mitigate cognitive biases as they do not rely on the human thought processes to quantify uncertainty. For instance, in Example 1, before releasing the drug for human tests, the drug must be taken through animal trials [49]. The trials done on animals could be relevant historical data, which could help formulate a prior distribution for human patients. Similarly, data from *Phase I* clinical trials can be used to formulate a prior distribution for *Phase II/III* clinical trials, and so on. It could be argued, however, that biases may have arisen in the creation of the historical model and data selection.

The disadvantage of these techniques is that they require access to similar research and/or historical data. Many subject fields may not have similar studies to the current research task, and/or access to historical data is restricted, rendering these methods futile. The problem of not having access to historical data is relevant to Example 2. Although it may be possible to find

information on a similar security failure in a similar organisation, the amount of data needed to form an appropriate distribution may be lacking. Also, as stated in Section 3.1.1, once a security threat has been observed, the system may be altered, making the new system of the similar organisation different to the original system in any potential historical data.

## 3.6 Eliciting Priors from Multiple Experts

The methods outlined in Sections 3.3-3.4 give an overview of appropriate techniques to elicit prior distributions from an individual expert. However, as outlined in Section 3.1.1, analysts should use multiple experts to form a probability distribution representative of the whole field of research [2, 3]. When selecting an appropriate method, the analyst must trade off the simplicity of the method and the likely accuracy of the resulting prior. Extensive reviews of methods for combining multiple expert beliefs can be found in [50, 51, 6] and [52]. A comparison of some aggregation techniques can be found in [3]. An overview of methods to combine information from multiple experts follows.

**Behavioural Aggregation Methods:** These methods involve groups of experts discussing their beliefs and, in some cases, reaching a consensus.

- **Delphi Method** [53]: This method most commonly starts with a questionnaire that is sent out to all the selected experts. Opinions from the initial questionnaire are then sent out to the same experts along with a new questionnaire, thus giving experts the option to update their own opinions based on their peers responses [54]. This process may be repeated until consensus is reached [50].

- **Interactive Group Methods**: In these methods, experts and a facili-

tator meet to discuss and come to a consensus on a probability distribution [50, 6]. The SHELF method [26] is a very common interactive group method. It is a framework for facilitators to follow that allows experts to reach a consensus.

Although behavioural aggregation may appear to be an ideal way to aggregate expert opinion because it allows for shared expertise, the goal of reaching a 'consensus' can be problematic. This is because social psychological factors can distort the process of reaching consensus [55]. Interactive groups are particularly vulnerable to groupthink, seniority, titles or social hierarchy. For example, people who are high on personality characteristics such as social dominance or who are considered by other experts in the group to have more expertise (e.g., more citations) can influence a consensus prior so that it resembles the dominant individual's prior and doesn't actually capture the variability in certainty across the group. This problem can be overcome in some variants of the Delphi method where experts are anonymous to each other (e.g., online methods).

Over-confidence may also be an issue for consensus priors from behavioural aggregation. In an evaluation of methods of aggregation, [3] compared the SHELF method with mathematical methods. They speculated that over-confidence may have been present in the SHELF method because it yielded narrower distributions. Moreover, not only is forming a consensus difficult, but the whole process of behavioural aggregation can take time [3]. For the interactive group technique, there can be difficulty in aligning work schedules, deciding on an optimal time to meet and the discussion itself. For the Delphi Method, there may be a significant wait time during the process of having to send responses back and forth to expert group members. All of these issues may be resolved through mathematical aggregation discussed below.

**Mathematical Aggregation Methods:** These methods involve mathe-

matical techniques to combine individual expert priors elicited by methods in the sections above.

- **Bayesian Method**: This approach treats each expert's prior as new data. The analyst updates their prior with the "new data". The resulting posterior consists of the combined prior beliefs [56, 57, 58]. This process can be complex and time consuming [6].

- **Opinion Pooling**: These methods form a consensus distribution $f(\theta)$ as a function of the individual distributions ($f_1(\theta)$, $f_2(\theta)$, ...) [6, 59]. A simple example of this is Linear Pooling where the consensus distribution is equal to the weighted sum of the individual distributions. That is,

$$f(\theta) = \sum_{i=1}^{n} w_i f_i(\theta) \tag{3.3}$$

  where $n$ is the number of experts. Other common pooling methods can be found in [6]. Key methods for finding weights are listed below:

  - Equal Weighting: Give each distribution the same weight, $1/n$.

  - Self-Weighting: Each expert gives themselves the appropriate weighting.

  - Expertise/Performance weighting: Not all experts may be considered equal. Some methods suggest weighting experts based on their expertise or performance in their field [51, 6, 60]. As an example, an analyst may decide that the longer someone has been working in the field, the higher the weight their prior receives. Cooke's Method [61], is a performance weighting method that gives the experts an added task (a smaller elicitation process where the subject mat-

ter parameter is known) to assess performance. An application of Cooke's method is outlined in [62].

– AHP Weighting: Also encountered in Section 3.3, an Analytic Hierarchy Process [33] is a process where an analyst can compare all solutions against a decision criterion [63]. In the case of weighting expert priors, [63] used three criteria for expert assessment: "years of experience", "number of observed failures", and "level of training". The weights obtained from the process are used in two ways: firstly, to create a shortlist of potential experts and secondly, as the weights for the aggregation process.

Selection of weights is a very important aspect of mathematical aggregation. Selection methods like equal weighting and self weighting are the easiest methods to apply. However, equal weighting treats all experts as being equally knowledgeable on the subject matter, which is generally not applicable to real life. Self-weighting methods can also be inaccurate. It has been found in the past that with self-weighting, women ranked their expertise lower than men [64]. This is still the case today where men tend to be more self-promoting than women [65]. Also, with overconfidence already being a bias present in expert elicitation, it may become particularly detrimental in this selection method. In contrast, methods based on performance, or some other objective criterion seem appropriate, as they take into account the level of expertise of any given expert in a less subjective manner. For an overview of selecting weights, in mathematical aggregation, see [50].

**Hybrid Methods:**

To utilise the benefits of behavioural aggregation and mitigate the drawbacks, there have been hybrid methods created that incorporate both behavioural and mathematical aggregation techniques [66]. One such protocol is the IDEA

protocol [67]. "It encourages experts to **I**nvestigate and estimate individual first round responses, **D**iscuss, **E**stimate second round responses, following which judgements are combined using mathematical **A**ggregation" [67].

## 3.7 Summary and Conclusion

When information from the likelihood is limited, obtaining an accurate and informative prior distribution is especially critical. This paper outlines some of the important approaches for eliciting prior distributions along with their merits and limitations. Each elicitation task is unique and so are the experts involved in that elicitation. Therefore, while it may be possible to find a suitable and appropriate prior elicitation approach for a given task, none of the approaches is arguably superior overall [25]. Moreover, there could be applications for which none of the existing methods are appropriate.

### 3.7.1 Persistent Challenges

Where expert involvement is required, there will always be persistent challenges in obtaining an informative prior. This paper does not address the issues surrounding selecting experts. However, experts need to be highly knowledgeable in the relevant content area, and for a number of the methods, they also need to both understand the relevant statistical concepts, and be able to think in a statistically valid manner.

Another related area of challenge is in minimising the influence of the common cognitive biases outlined in Section 3.2. They are part of natural human thought processes that have proved adaptive in other contexts, particularly when making decisions under pressure. It is not realistic to expect that they can be eliminated. But a key goal of any expert elicitation method should be to reduce them as much as possible.

Recall Example 2, where the task is to elicit a prior distribution using related information $x$ that could be heterogeneous both in type and relevance. Common methods discussed in Section 3.3 and 3.4 look at getting information solely from an expert, methods in Section 3.5 look at solely getting information from the data. However, methods that enable the analyst to elicit a prior distribution using all of the heterogeneous relevant information available, as well as including expert knowledge, do not yet exist.

Translating the uncertainty of an expert into a probability distribution may always remain a challenging problem. As discussed, there are many ways to do this. However, the most suitable method is usually task specific and should be thought of as such. There is no one superior way to quantify someone's uncertainty. The process used for each elicitation task should be selected based on the problem at hand, the resources available, the experts available, and their ability to quantify their uncertainty, as well as, the availability of any past research/data or relevant information.

It is possible that an elicited prior distribution may not accurately capture the expert's beliefs and can be considered to be only an approximation at best. Importantly, it may never be possible to ascertain how accurately a prior distribution reflects an expert's beliefs. Inaccurate elicitation of prior distributions may lead to inaccurate posterior inference and therefore, to inaccurate data analysis. This problem is exacerbated when the prior dominates the likelihood because of insufficient data.

A possible solution to inaccurately elicited prior distributions is to implement a prior robustness analysis. This typically involves defining a class of prior distributions that encompass the uncertainty around the original prior [68]. Prior robustness analysis studies the sensitivity of the posterior distribution to the choice of prior distribution. Several prior robustness approaches have been developed; see [69] for a general introduction to prior robustness analysis and an overview of various approaches therein. A recent approach to implement prior robustness, which is also straightforward to implement, uses distortion

functions [70] to form a *distorted band* class of priors. One of the challenges in implementing a prior robustness analysis is in quantifying the uncertainty in the elicitation of the original prior. Methods to quantify such uncertainty have also been proposed. For the distorted band class of priors, [71] suggests a simple interrogation approach that can determine the length of the distorted band class of priors to accurately quantify the expert's uncertainty around the original prior distribution.

In this manuscript, the focus is on how an informative prior can be elicited, that is, capturing an expert's beliefs to form a probability distribution. This should not be confused with obtaining forecasts or estimates from an expert, which although can contain aspects of what is discussed in this paper, is not forming a probability distribution from their beliefs. For approaches of this nature, we suggest the reader look into the "Good Judgement Project" and the research related to it [72, 73, 74].

### 3.7.2 Future Research

Persistent challenges open the way for research into new methods. Suggestions for potential research avenues are as follows.

There is a need to find the most appropriate method for a given task. Therefore, the field of prior elicitation could benefit from having more research into comparisons of different methods. Although some research on comparing different methods is available, it often focuses only on one type of elicitation technique, see for example [25, 32], or [3]. More research into comparisons of a range of approaches will not only help practitioners gauge which method is more appropriate for their task, but it will allow for continual discussion on the topic of elicitation.

Research could also consider situations similar to the specific case of Example 2, where data for the likelihood is unlikely to occur and when it does occur, it leads to system change. This implies that the initial prior needs to be reliable

and, may need to be updated as the system changes. Redoing the same initial elicitation process could be a solution, but this process will be time consuming and not always a viable option, so, other methods should be explored. Additionally, how to incorporate all the heterogeneous information available for the elicitation, would be an interesting topic for exploration.

New prior elicitation techniques that focus on addressing some of the key challenges should also be researched further. One such technique could be on producing priors from modelling the expert decision making process. There do exist methods that model a simple decision making task, such as the graphical elicitation method discussed in Section 3.4, and also methods briefly considered in Section 3.3.2 (e.g. the ranking method). However, these methods rely on hypothetical decision making; that is, making decisions on circumstances that are not real. Research could instead focus on eliciting priors from real life decision making.

Using real situations could lead to a more realistic prior than a hypothetical decision making task, because decisions made in real life carry far greater consequences than those in hypothetical situations, leading experts to make more effort to be accurate in the information they provide. It is worth noting that this decision making task may not be performed under the scope of prior elicitation; there may be past decisions that have been made that can be used to elicit priors from the decision makers. Using past decisions could help in creating a calibration technique, like those discussed in Section 3.2, as the data may include the real life outcome after the expert's decision. This could allow for the analyst to compare the decision against the actual outcome and adjust for any biases that may arise in the decision making process, resulting in a fair prior. Another benefit of such a method is that the expert requires no statistical knowledge to elicit a prior, which can be a drawback of other methods.

The role of prior robustness in prior elicitation should also be explored further. Focus could be directed towards the expert's uncertainty surrounding

their prior and using different approaches to obtain this uncertainty to form bounds on the elicited prior distribution. One way could be to use a performance based criterion, like those discussed when combining multiple expert priors (Section 3.6), in which experts who are considered "more knowledgeable" in the field will have narrower bounds on their elicited prior.

# References

[1] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974.

[2] Anthony O'Hagan. Expert knowledge elicitation: subjective but scientific. *The American Statistician*, 73(sup1):69–81, 2019.

[3] Cameron J Williams, Kevin J Wilson, and Nina Wilson. A comparison of prior elicitation aggregation using the classical method and shelf. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2021.

[4] Anca M Hanea, Victoria Hemming, and Gabriela F Nane. Uncertainty quantification with experts: present status and research needs. *Risk Analysis*, 2021.

[5] Lionel A Galway. Subjective probability distribution elicitation in cost risk analysis: A review. 2007.

[6] Anthony O'Hagan, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons, 2006.

[7] David Jenkinson. The elicitation of probabilities: A review of the statistical literature. Technical report, Citeseer, 2005.

[8] Angelika M Stefan, Nathan J Evans, and Eric-Jan Wagenmakers. Practical challenges and methodological flexibility in prior elicitation. *Psychological Methods*, 2020.

[9] R Hogarth. Judgement and choice wiley. *New York*, 1987.

[10] Paul H Garthwaite, Joseph B Kadane, and Anthony O'Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–701, 2005.

[11] Jack B Soll and Joshua Klayman. Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2):299, 2004.

[12] Joshua Klayman, Jack B Soll, Claudia Gonzalez-Vallejo, and Sema Barlas. Overconfidence: It depends on how, what, and whom you ask. *Organizational behavior and human decision processes*, 79(3):216–247, 1999.

[13] Allen Parducci. Range-frequency compromise in judgment. *Psychological Monographs: General and Applied*, 77(2):1, 1963.

[14] Christoph Werner, Tim Bedford, Roger M Cooke, Anca M Hanea, and Oswaldo Morales-Napoles. Expert judgement for dependence in probabilistic modelling: A systematic literature review and future research directions. *European Journal of Operational Research*, 258(3):801–819, 2017.

[15] Martine J Barons, Steven Mascaro, and Anca M Hanea. Balancing the elicitation burden and the richness of expert input when quantifying discrete Bayesian networks. *Risk Analysis*, 2021.

[16] Mary A Meyer and Jane M Booker. *Eliciting and analyzing expert judgment: a practical guide*. SIAM, 2001.

[17] DV Lindley. The improvement of probability judgements. *Journal of the Royal Statistical Society: Series A (General)*, 145(1):117–126, 1982.

[18] Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D Phillips. Calibration of probabilities: The state of the art. *Decision making and change in human affairs*, pages 275–324, 1977.

[19] Tommi Perälä, Jarno Vanhatalo, Anna Chrysafi, et al. Calibrating expert assessments using hierarchical gaussian process models. *Bayesian analysis*, 15(4):1251–1280, 2020.

[20] J Edward Russo and Paul JH Schoemaker. Managing overconfidence. *Sloan management review*, 33(2):7–17, 1992.

[21] GR Chesley. Elicitation of subjective probabilities: A review. *The Accounting Review*, 50(2):325–337, 1975.

[22] Sindhu R Johnson, George A Tomlinson, Gillian A Hawker, John T Granton, and Brian M Feldman. Methods to elicit beliefs for Bayesian priors: a systematic review. *Journal of clinical epidemiology*, 63(4):355–369, 2010.

[23] Robert L Winkler. The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical association*, 62(319):776–800, 1967.

[24] Carl S Spetzler and CS Stael Von Holstein. *Probability encoding in decision analysis*. Stanford Res. Inst., 1972.

[25] Joseph Kadane and Lara J Wolfson. Experiences in elicitation: [read before the royal statistical society at a meeting on'elicitation 'on wednesday, april 16th, 1997, the president, professor afm smith in the chair]. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):3–19, 1998.

[26] Jeremy E Oakley and Anthony O'Hagan. Shelf: the sheffield elicitation framework (version 2.0). *School of Mathematics and Statistics, University of Sheffield, UK (http://tonyohagan. co. uk/shelf)*, 2010.

[27] David E Morris, Jeremy E Oakley, and John A Crowe. A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software*, 52:1–4, 2014.

[28] Joseph B Kadane, James M Dickey, Robert L Winkler, Wayne S Smith, and Stephen C Peters. Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association*, 75(372):845–854, 1980.

[29] Artur Akbarov et al. *Probability elicitation: Predictive approach*. PhD thesis, University of Salford, 2009.

[30] Marcelo Hartmann, Georgi Agiashvili, Paul Bürkner, and Arto Klami. Flexible prior elicitation via the prior predictive distribution. In *Conference on Uncertainty in Artificial Intelligence*, pages 1129–1138. PMLR, 2020.

[31] Robert L Winkler. The quantification of judgment: Some methodological suggestions. *Journal of the American Statistical Association*, 62(320):1105–1120, 1967.

[32] Ali E Abbas, David V Budescu, Hsiu-Ting Yu, and Ryan Haggerty. A comparison of two probability encoding methods: Fixed probability vs. fixed variable values. *Decision Analysis*, 5(4):190–202, 2008.

[33] Roseanna W Saaty. The analytic hierarchy process—what it is and how it is used. *Mathematical modelling*, 9(3-5):161–176, 1987.

[34] TL Saaty. The analytic hierarchy process mcgraw hill, new york. *Agricultural Economics Review*, 70, 1980.

[35] Enrico Cagno, Franco Caron, Mauro Mancini, and Fabrizio Ruggeri. Using ahp in determining the prior distributions on gas pipeline failures in a robust Bayesian approach. *Reliability Engineering & System Safety*, 67(3):275–284, 2000.

[36] Robert T Eckenrode. Weighting multiple criteria. *Management science*, 12(3):180–192, 1965.

[37] Ward Edwards and F Hutton Barron. Smarts and smarter: Improved simple methods for multiattribute utility measurement. *Organizational behavior and human decision processes*, 60(3):306–325, 1994.

[38] Chen Wang and Vicki M Bier. Expert elicitation of adversary preferences using ordinal judgments. *Operations Research*, 61(2):372–385, 2013.

[39] Johannes G Jaspersen and Gilberto Montibeller. Probability elicitation under severe time pressure: A rank-based method. *Risk Analysis*, 35(7):1317–1335, 2015.

[40] Sheila M Gore. Biostatistics and the medical research council. *Medical Research Council News*, 35:19–20, 1987.

[41] Ivan R Diamond, Robert C Grant, Brian M Feldman, George A Tomlinson, Paul B Pencharz, Simon C Ling, Aideen M Moore, and Paul W Wales. Expert beliefs regarding novel lipid-based approaches to pediatric intestinal failure–associated liver disease. *Journal of Parenteral and Enteral Nutrition*, 38(6):702–710, 2014.

[42] Christopher J Casement and David J Kahle. Graphical prior elicitation in univariate models. *Communications in Statistics-Simulation and Computation*, 47(10):2906–2924, 2018.

[43] Geoffrey Jones and Wesley O Johnson. Prior elicitation: Interactive spreadsheet graphics with sliders can be fun, and informative. *The American Statistician*, 68(1):42–51, 2014.

[44] Owen Thomas, Henri Pesonen, and Jukka Corander. Probabilistic elicitation of expert knowledge through assessment of computer simulations. *arXiv preprint arXiv:2002.10902*, 2020.

[45] S James Press. *Subjective and objective Bayesian statistics: Principles, models, and applications*, volume 590. John Wiley & Sons, 2009.

[46] Joseph G Ibrahim, Ming-Hui Chen, et al. Power prior distributions for regression models. *Statistical Science*, 15(1):46–60, 2000.

[47] Joseph G Ibrahim, Ming-Hui Chen, Yeongjin Gwon, and Fang Chen. The power prior: theory and applications. *Statistics in medicine*, 34(28):3724–3749, 2015.

[48] Ming-Hui Chen, Amita K Manatunga, and Christopher J Williams. Heritability estimates from human twin data by incorporating historical prior information. *Biometrics*, pages 1348–1362, 1998.

[49] J Rick Turner. *New drug development: an introduction to clinical trials*. Springer Science & Business Media, 2010.

[50] Jonathan Rougier, Lisa J Hill, and Robert Stephen John Sparks. *Risk and uncertainty assessment for natural hazards*. Cambridge University Press, 2013.

[51] Robert L Winkler. The consensus of subjective probability distributions. *Management science*, 15(2):B–61, 1968.

[52] Robert T Clemen and Robert L Winkler. Combining probability distributions from experts in risk analysis. *Risk analysis*, 19(2):187–203, 1999.

[53] Olaf Helmer. Analysis of the future: The delphi method. Technical report, Rand Corp Santa Monica CA, 1967.

[54] Marlen Niederberger and Julia Spranger. Delphi technique in health sciences: A map. *Frontiers in public health*, 8:457, 2020.

[55] Robert S Baron. So right it's wrong: Groupthink and the ubiquitous nature of polarized group decision making. *Advances in experimental social psychology*, 37(2):219–253, 2005.

[56] Peter A Morris. Decision analysis expert use. *Management Science*, 20(9):1233–1241, 1974.

[57] Dennis Lindley. Reconciliation of probability distributions. *Operations Research*, 31(5):866–880, 1983.

[58] Isabelle Albert, Sophie Donnet, Chantal Guihenneuc-Jouyaux, Samantha Low-Choy, Kerrie Mengersen, and Judith Rousseau. Combining expert opinions in prior elicitation. *Bayesian Analysis*, 7(3):503–532, 2012.

[59] Christian Genest and James V Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, 1986.

[60] Dennis V Lindley and Nozer D Singpurwalla. Reliability (and fault tree) analysis using expert opinions. *Journal of the American Statistical Association*, 81(393):87–90, 1986.

[61] Roger Cooke et al. *Experts in uncertainty: opinion and subjective probability in science*. Oxford University Press on Demand, 1991.

[62] Willy Aspinall. A route to more tractable expert advice. *Nature*, 463(7279):294–295, 2010.

[63] Zaki Syed, Oleg Shabarchin, and Yuri Lawryshyn. A novel tool for Bayesian reliability analysis using ahp as a framework for prior elicitation. *Journal of Loss Prevention in the Process Industries*, 64:104024, 2020.

[64] Klaus Brockhoff. *IV. E. The Performance of Forecasting Groups in Computer Dialogue and Face-to-face Discussion*, volume 68. Addison Wesley Publishing Company, 1975.

[65] Christine L Exley and Judd B Kessler. The gender gap in self-promotion. Technical report, National Bureau of Economic Research, 2019.

[66] William R Ferrell. Discrete subjective probabilities and decision analysis: Elicitation, calibration and combination. 1994.

[67] AM Hanea, MF McBride, MA Burgman, and BC Wintle. Classical meets modern in the idea protocol for structured expert judgement. *Journal of Risk Research*, 21(4):417–433, 2018.

[68] Sanjib Basu. Variations of posterior expectations for symmetric unimodal priors in a distribution band. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 320–334, 1994.

[69] D. Ríos Insua and F. Ruggeri. *Robust Bayesian Analysis*. Springer, New York, 2000.

[70] J Pablo Arias-Nicolás, Fabrizio Ruggeri, and Alfonso Suárez-Llorens. New classes of priors based on stochastic orders and distortion functions. *Bayesian Analysis*, 11(4):1107–1136, 2016.

[71] Chaitanya Joshi, Fabrizio Ruggeri, and Simon P Wilson. Prior robustness for Bayesian implementation of the fault tree analysis. *IEEE Transactions on Reliability*, 67(1):170–183, 2018.

[72] Philip E Tetlock and Dan Gardner. *Superforecasting: The art and science of prediction*. Random House, 2016.

[73] Lyle Ungar, Barbara Mellers, Ville Satopää, Philip Tetlock, and Jon Baron. The good judgment project: A large scale test of different methods of combining expert predictions. In *2012 AAAI Fall Symposium Series*, 2012.

[74] Barbara Mellers, Lyle Ungar, Jonathan Baron, Jaime Ramos, Burcu Gurcay, Katrina Fincher, Sydney E Scott, Don Moore, Pavel Atanasov, Samuel A Swift, et al. Psychological strategies for winning a geopolitical forecasting tournament. *Psychological science*, 25(5):1106–1115, 2014.

# Chapter 4

# Eliciting Informative Priors by Modelling Expert Decision Making

There are significant limitations to current methods for eliciting the prior beliefs of experts. To combat some of these limitations, this paper proposes an alternative approach that infers an expert's prior beliefs about an uncertain event, $A$, from the expert's past decisions. We show that an analyst can use past information on an expert's decision-making task, contingent on an expert's prior of $A$, to model the decision-making process and infer an approximation of the prior for $A$. This concept is illustrated by an application to recidivism. We conclude this work by highlighting important directions for future research.

## 4.1 Introduction

Beginning with some prior knowledge (a prior probability distribution), Bayesian inference updates the prior by taking information from observed data (a likelihood) to build a posterior distribution over the parameters of interest, $\theta$:

$$p(\theta|y) \propto p(\theta)p(y|\theta), \tag{4.1}$$

A prior distribution that has minimal influence on the posterior distribution, a 'non-informative' prior, is often used. Where there is large amounts of data, the choice of prior is largely irrelevant since the likelihood dominates the posterior distribution. However, if there is limited data, the influence from the likelihood becomes minimal, producing a posterior that relies heavily on the prior information. For such instances, an informative prior distribution could be used [1].

Table 4.1: Definitions Expanded from a Table in [2]

| **Name** | **Description** |
|---|---|
| *Prior Elicitation* | The process of obtaining knowledge from a source to form a prior distribution that can be used for further Bayesian analysis. |
| *Expert* | An individual (or a group of individuals) who has extensive knowledge on a certain subject matter. The expert is also referred to as the decision maker in this text. |
| *Decision Maker* | The individual who performs a decision making task. In most cases, the Decision Maker and the Expert will be the same individual. |
| *Analyst* | An individual who performs the task of forming a prior distribution using prior elicitation techniques. |
| *Facilitator* | An individual who performs the task of eliciting knowledge. In some cases, the Facilitator and the Analyst may be the same individual. |

We consider scenarios exhibiting an event, $A$, that is of serious consequence and where data on $A$ is limited as the event rarely occurs. An analyst (see Table 4.1 for definitions used throughout this paper) wishes to obtain an informative prior distribution for $A$. Although there may not be any data on $A$, there may be some other related source of information that can be used to obtain a prior for $A$. The most common way to do this is to elicit a distribution from an expert in the relevant field of interest. Methods to obtain an informative prior distribution from an expert are described in [2], which assigns methods to three categories; 1) Direct Interrogation Methods, 2) Indirect Interrogation Methods, and 3) Graphical/ Visual Methods. Direct Interrogation methods

[3, 4, 5] involve asking experts about the probability distributions directly. This can be challenging because experts must first have a firm grasp of probability theory and distributions. There are circumstances where an expert can first be taught key probability concepts [3, 6, 7], but this can prove difficult and create inaccurate prior distributions [8, 9]. This issue can also be seen in some graphical/visual methods [2]. Indirect Interrogation methods have been introduced to help combat the requirement of experts needing knowledge of probability theory. Indirect Interrogation methods involve asking the expert questions that are not directly based on the probability distributions themselves, but instead are easy for the expert to comprehend. From there, an analyst will use mathematical logic to infer a prior distribution. Two examples of Indirect Interrogation that display the simplicity of questioning are: getting the expert to place bets on which event they think is more likely [10] and getting the expert to rank the likelihood of events [11, 12, 9]. As highlighted in [2], some Indirect Interrogation methods can be thought of as hypothetical decision-making tasks. Hypothetical decision-making implies that whether the decision is correct or incorrect has no real consequence for the expert. Therefore, prior elicited in this way may not accurately reflect the expert's thinking in real life.

The use of experts during the process of elicitation has the added complexity of introducing cognitive and motivational biases. In Direct Interrogation elicitation, the simple mistake of asking a question a certain way can produce cognitive biases which influence the experts response (e.g., anchoring and adjusting [13], where values in the questions are used by an expert to anchor their response value). Prior elicitation methods that use experts may also have cognitive biases based on an expert's work experience (e.g., judgment by availability [13], where an expert will put more weight on an event just because the expert witnessed that event more recently) or, to put it more generally, an expert's life experience, that includes biases they have formed over time (e.g.,

gender bias, racial bias). Using a group of experts to elicit one prior distribution can help an analyst gain a wider view of the whole field of interest [14]. A common way to do this is to get a group of experts to discuss opinions to form a consensus, however, this method can also come with cognitive biases that an analyst should be aware of, such as *groupthink* [15]. Groupthink is where the need to reach a consensus, while maintaining harmony within the group, means individuals do not voice alternative perspectives that may be outside the social "norm" or maybe against the perspective of a strongly influential individual, skewing the group's elicited prior in one direction [15]. Instead of having experts reach a consensus, some methods allow analysts to combine experts' individual priors mathematically [3] to avoid cognitive biases that are formed from group consensus, such as groupthink. Some methods can elicit a prior distribution without an expert's input by using historical data (e.g., use the posterior from a similar historical study [16]), however, in most cases this historical data will not exist. Also, historical data is not immune from the effects of biases, and it is not just an individual expert's cognitive biases that an analyst must be aware of. Sometimes available data might encompass societal biases [17]. A famous example is the Correctional Offender Management Profiling for Alternative Sanction, COMPAS [18]. COMPAS was a risk assessment tool that was used to obtain a recidivism score for defendants. Although ethnicity was not a factor in the model, the tool was still more likely to class black individuals as high risk than other individuals [18]. This was because the model had learned from historic discriminatory court cases and enhanced the prejudices in the judicial system [17]. Another example is a tool that was used to rank the top five applicants based on their resumes for job vacancies at Amazon; it was found to be penalising applicants that were women and favouring those that were male [19, 17]. This was because the model learned patterns from historic data where women were not hired for positions at tech companies [17]. The societal biases of blacks being more likely to commit crimes and females being less adequate for specific jobs were

shown in the data applied to these models and influenced the outputs. Lack of information or inadequate information can also produce a bias [20]. If the available information is heavily dominated by information on one group then it is obvious that the results produced with this information could be considered biased, like in the COMPAS example. Often the available information is tabular data, which may be missing key information that is needed to give accurate outputs. Tabular data variables may also represent multiple factors of interest that are not directly collected in the data (confounding variables), making it hard to understand what variables are truly influencing the output. Reducing the impact of biases on the elicited prior is a key interest in prior elicitation [14, 3].

### 4.1.1 Motivation

We believe the key limitations of current methods are: a) the statistical knowledge required of experts to perform elicitation by Direct Interrogation methods, b) the "hypothetical" decision-making tasks in Indirect Interrogation methods that have no real-life impact and could affect the accuracy of the elicited prior, and, c) the difficulty of identifying biases when eliciting an expert prior. We introduce a concept that eliminates some of these limitations by eliciting an approximation of a prior distribution through modelling an expert's past decision-making tasks. Our method eliminates the statistical knowledge required by utilising a decision-making task that an expert performs as part of their duties. Often this decision-making task has real-life implications, meaning more importance is placed on the decision, and the experts will strive to be more accurate in their decisions. Thus, by modelling their past decisions, we may be able to capture their thinking more accurately than methods that rely on hypothetical decision-making. Also, modelling past real-life decisions eliminates biases that could be introduced in direct interrogation methods. Although, because we are using experts, there may still be cognitive biases af-

fecting the elicited distribution. Modelling data from the past decision-making tasks may allow analysts to identify variables that may be considered to be inducing bias in the decision-making process. Our goal is to introduce the broader concept to the reader and provide a simple example that highlights the use of this concept. The method is explained further in Section 4.2. We discuss ways to assess model behaviour in Section 4.3, with Section 4.4 outlining a simple example application. Finally, we close in Section 4.5 with a summary of conclusions and further work.

## 4.2 Eliciting Uncertainty from Decision Making

We introduce a method that combines concepts from Indirect Interrogation methods as well as those that use historical data, by forming a prior distribution from an expert's past decision-making task. We are concerned with an undesirable future event $A$. The expert wishes to prevent $A$ from occurring and considers a (preventative) decision $Y$. Let $X$ be the information that is available to the decision maker at the time. The expert is interested in being able to quantify the prior probability on $A|X$, that is, what is the probability that $A$ will occur given the available information $X$. Using the expert's past decisions, the decision process $Y|X$ can be modelled. We conjecture that given $X$, the uncertainty in the outcome of $Y$ reflects the experts' uncertainty on whether $A$ would occur or not if no preventative measures were taken. Therefore, $A|X$ and $Y|X$ are intimately related. For simplicity, we assume that the event $A$ is binary (*occurs or not*), so also is the preventative decision $Y$ (*prevention put in place or not*). Let $Y|X \sim Bernoulli(p)$ and $A|X \sim Bernoulli(q)$. To be able to model the decision-making process $Y|X$ accurately, the process should be repetitive (carried out often) and its outcomes and the information used to make the decision should be available.

Let $Y_i$ denote the decision made at the $i^{th}$ instance (hereafter referred to as a *case*) and $X_i$ be the information used by the decision maker to make that decision. Suppose that the data on $n$ cases is available so that we have $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_n\}$ and $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$. Let $\boldsymbol{\theta}$ be model parameters that link the decisions $\mathbf{Y}$ to the available information $\mathbf{X}$ such that $\mathbf{Y} \sim f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})$. Given, $\mathbf{Y}$, $\mathbf{X}$ and a prior distribution $\pi(\boldsymbol{\theta})$, we can find the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{X})$. Assuming information on a sufficient number $n$ of similar cases and an appropriate model $f$, it is reasonable to believe that using the information $X^*$ for the next case, we could accurately predict the decision $Y^*$ that the decision maker is likely to make using the posterior predictive distribution.

$$P(Y^*|X^*, \mathbf{Y}, \mathbf{X}) = \int P(Y^*|X^*, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{X})\, d\boldsymbol{\theta}. \tag{4.2}$$

Let $A_i$ be the undesirable consequence for the $i^{th}$ case, which may or may not materialize. The data on (some or all of) past $A_i$ may be available, but that is not considered here at this stage. Since $Y_i$ is the preventative decision to mitigate the risk of $A_i$, it is clear that $Y_i$ reflects the decision maker's prediction on $A_i$. That is, that a preventative decision was put in place implies that the decision maker believes that $A_i$ is likely to occur. Similarly, if the preventative measures were not put in place, this would reflect the decision maker's belief that $A_i$ is unlikely to occur. That is,

$$A_i|X_i \overset{\text{d}}{\approx} Y_i|X_i. \tag{4.3}$$

Therefore, given the information $X^*$ for the next case, the conditional predictive prior for $A^*$ can be approximated using the posterior predictive distribution in Equation (4.2). That is,

$$\pi(A^*|X^*) \approx P(Y^*|X^*, \mathbf{Y}, \mathbf{X}). \tag{4.4}$$

See the accompanying influence diagram (Figure 4.1) that depicts the relationship between the variables.

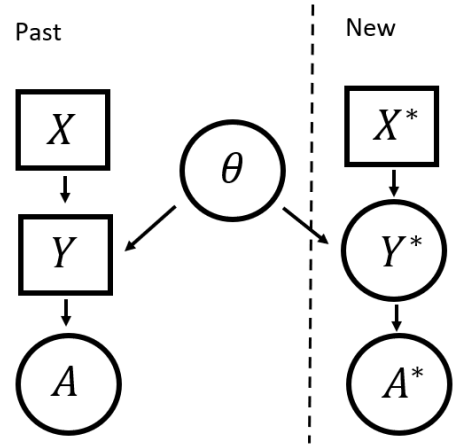



Figure 4.1: Influence Diagram for Eliciting Prior Distributions from Expert Decision Making

As an illustrative example, let $A$ be the event that a property in an industrial area will be burgled. This threat could be potentially mitigated by employing the services of a security consultant who would review the relevant information, $X$, make an assessment, $Y$, about the imminent risk and provide recommendations of security features that could be installed to prevent the threat from eventuating. If the data on $n$ recent property evaluations by the same consultant are available, then we can model the consultant's risk perception using a statistical model. The goal is to obtain the probability distribution of a new property being burgled using the relevant information available $X$. This probability distribution can be considered as an approximation to the consultant's prior probability distribution on whether the event $A$ will occur given $X$.

Note that our goal is not to accurately predict $A$. Instead, we want the model to accurately mimic the experts' decision-making process, and capture the experts' uncertainty about the event $A$, by considering the uncertainty in the model for the surrogate event $Y$. To ascertain whether the model is accu-

rately mimicking the expert's decision-making process, an analyst can observe at least one of the measures of central tendency of the elicited probability distribution and assess whether it correctly predicts $Y_i$ in most of the cases (see Section 4.3). Moreover, we conjecture that the aleatory uncertainty captured by the model reflects the aleatory uncertainty of the expert on whether $A$ will occur or not given $X$. Our conjecture assumes that the decision maker recognizes that due to natural variability, an event may or may not occur even when it is very likely to occur and vice versa.

While this approach doesn't mandate an expert to have sufficient knowledge of statistics, it puts a heavier burden on an analyst's statistical skills. This is because the analyst must be capable of precisely modelling the decision-making process. Choosing and refining the models will necessitate a solid statistical foundation. We will illustrate the use of this method with an example in Section 4.4 using Bayesian logistic regression. Given $Y_i|X_i \sim Bernoulli(p_i)$, the logistic regression model, with a link function $g(.)$, is represented as,

$$g(p_i) = \theta_0 + \theta_1 x_{1i} + ...$$

For example, with a simple logit link function and one predictor variable,

$$logit(p_i) = log(\frac{p_i}{1-p_i}) = \theta_0 + \theta_1 x_i$$

$$\Rightarrow p_i = \frac{exp(\theta_0 + \theta_1 x_i)}{1 + exp(\theta_0 + \theta_1 x_i)} \tag{4.5}$$

A Bayesian approach is implemented by placing prior distributions on the model parameters, $\boldsymbol{\theta} = \{\theta_0, \theta_1, ...\}$. Sampling methods, such as MCMC methods, can be used to approximate the posterior distribution of $\boldsymbol{\theta}$. An analyst can select the prior distribution for model parameters and the sampling method

and adjust them to build the most appropriate model (Section 4.3). To approximate the probability distribution for $p_i$ from this model, we can sample from the posteriors of the model parameters, $\pi(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{X})$. These samples will be used in the model equation (for example Equation 4.5) to obtain samples of $p_i$. An approach such as the methods of moments can then be used to fit a Beta distribution to these samples, which forms the elicited prior distribution of $q_i$ for the model $A_i|X_i \sim Bernoulli(q_i)$.

There are many models that are used to predict rare or undesirable events, including Bayesian logistic regression models (e.g., for predicting recidivism [21, 22, 23, 24]). However, these models, to the best of our knowledge, have not yet been used to model expert decision-making or, to elicit an experts' prior distributions. We reiterate that our goal is not to predict a rare or undesirable event, instead, we wish to capture the uncertainty surrounding said event occurring.

## 4.3 Model Selection Diagnostics

To be able to elicit expert uncertainty accurately, we expect our model to behave like a decision-maker. We want it to be more uncertain when it sees data it has never seen before (wider distributions of $p_i$ that could be centered around 0.5) and less uncertain when it encounters familiar data (narrower distributions). Looking at the accuracy of the model is standard practice when assessing model performance (how accurately the model is predicting the response variable, $Y$, for a given test data set). If we wish to obtain the accuracy of a model which predicts the probability, $p_i$, of a binary decision, $Y_i$, labels are typically assigned as follows. If $p_i$ is less than 0.5 then the decision is labeled "no" and if $p_i$ is greater than 0.5 then the decision is labelled "yes" (or whatever the labels may be). When we are taking samples of $p_i$, it is common

practice to take the mean of those samples as our estimate of $p_i$ that assigns labels. However, to assess how well the model captures the experts' thinking, model accuracy is not the only diagnostic that is of importance, as we must also take into consideration the variability of the elicited distributions and the uncertainty that they capture.

It is easy to show that using the mean, of the sampled $p_i$ values, to assign la-

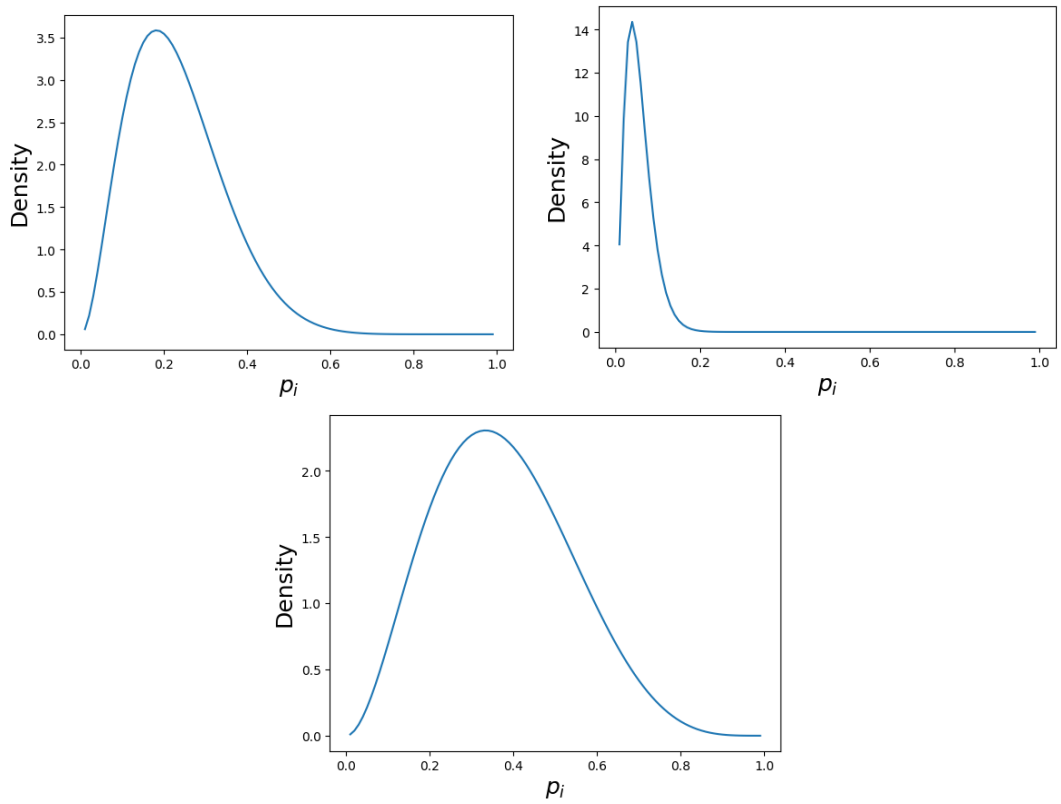



Figure 4.2: Distributions of $p_i$ for three Individuals that would obtain the same label assigned based on mean probability prediction.

bels for model accuracy may not give a fair representation of the variability of the distributions. For example, Figure 4.2 shows three distributions where the model would assign the same label if the means of $p_i$ were used for assigning labels. However, we can see that using the mean does not accurately capture the difference in variability of the distributions and that using the median or the mode of the distribution of $p_i$ would have assigned labels differently. We could also gain further insights by looking at the credible intervals of the distributions and assigning labels on whether the value of $p_i$ needed to assign a

certain label, lies within the credible interval. The credible interval also allows an analyst to assess the uncertainty of the elicited distributions, which is of importance when selecting an appropriate model. If the credible interval is wide and contains 0.5, then we can assume that our expert is fairly uncertain, and if it is narrow and on either side of 0.5, we can assume they are fairly certain. In the same way, the area *area under curve* (AUC) of the distribution can be used. To further assess the model's capabilities to capture uncertainty, an analyst can observe the entropy of the elicited distributions. If the entropy value is close to zero, then we assume the expert is fairly certain; if it is close to one, then we assume they are fairly uncertain. Assessing whether or not the model is behaving appropriately is case specific. If the analyst knows the decision making task has a lot of uncertainty, then they would expect high entropy values and will need to assess the trade-off between high entropy and high accuracy values. However, if the task is fairly certain, involving black and white responses, then we would expect low entropy values and aim for high accuracy from our model.

Table 4.2: Model Diagnostics we Suggest to Help Select an Appropriate Model for Prior Elicitation.

.

| **Name** | **Description** |
|---|---|
| *Mean Accuracy* | Percentage of correct predictions the model makes by using the mean of the sampled probabilities $p_i$. |
| *Mode Accuracy* | Percentage of correct predictions the model makes by using the mode of the sampled probabilities $p_i$. |
| *Median Accuracy* | Percentage of correct predictions the model makes by using the median of the sampled probabilities $p_i$. |
| *Area Under Curve (AUC) Accuracy* | Percentage of correct predictions the model makes by taking the largest area either side of 0.5 as the measure to form the model prediction. |
| *95% Credible Interval (CI) Accuracy* | Percentage of correct predictions the model makes by observing the 95% CI of $p_i$. If the 95% CI contains 0.5 then the assigned label can be either "Accept" or "Reject" and is a correct prediction. If the 95% CI is contained below 0.5 and the true label is "Accept" then it is a correct prediction. If the 95% CI is contained above 0.5 and the true label is "Reject" then it is a correct prediction. |
| *Percentage of the 95% CI correct predictions that contain 0.5.* | This will allow the analysts to see how many central distributions are elicited. |
| *Percentage of the 95% CI correct predictions that are either side of 0.5.* | This will allow the analysts to see how many skewed distributions are elicited. |
| *F-Score [25]* | A measure which shows the specificity (true negative rate) and sensitivity (true positive rate) of the model. The mean of the samples of $p_i$ is used to assign labels. The highest possible value of an F-score is 1.0, indicating perfect specificity and sensitivity, and the lowest possible value is 0, if either the specificity or the sensitivity is zero. $$F = 2\frac{specificity \times sensitivity}{specificity + sensitivity}$$ . |
| *Confusion Matrix [26]* | Shows the percentage of the mean predictions by whether the prediction is a true negative, true positive, false negative or false positive, showing the specificity and sensitivity of the model. The mean of the samples of $p_i$ is used to assign labels. |
| *Entropy [27]* | A measure of the amount of uncertainty in a distribution. A narrow distribution will give a value close to zero (showing a certain prediction), and a wide distribution will give a value close to 1 (showing an uncertain distribution). To make sure the model is behaving correctly, it will be helpful to observe a histogram of all entropy values for the training set, as well as observing the histograms of the entropy values of correct and incorrect predictions separately. |
| *Calibration Plot* | A calibration plot shows how well the prediction probabilities match the true percentage probabilities of the data. The mean of the samples of $p_i$ is used as prediction probabilities. |

These suggested diagnostics help an analyst assess the performance of the model, without looking at every single distribution produced. We advise analysts to look at multiple different model diagnostics to make sure the model is suitable for the task of prior elicitation and, also, to ensure they have a well-fitted model (Table 5.1). The analyst's goal should be to maximise the model's accuracy (how well it is predicting the response for a given data set) while also producing distributions that accurately capture uncertainty.

## 4.4 Example

Let $A$ be the event that a prisoner commits a crime upon release from prison. Information on a specific prisoner re-offending is limited and often censored, as we only know if a released prisoner commits a crime if they were caught. However, there exists an expert decision-making process that can be used to infer a prior distribution on the event $A$. This is the parole board hearing process. The parole board considers a report from a prisoner's case worker and decides whether or not to give a prisoner parole. When making a decision, the parole board is already taking into consideration the risk of the prisoner re-offending upon release, so this decision-making process can be used to infer a prior distribution on $A$. For example, if parole is not granted, this implies that the risk of re-committing a crime for an individual is high.

### 4.4.1 Data

We use a publicly available data set from the New York State Parole Board's interview calendar made available by The Parole Hearing Data Project [1]. This data set contains information on the prisoner, the hearing process, and the

[1]Data source `https://github.com/rcackerman/parole-hearing-data`

final decision[2]. It has 46 variables in total. We choose to take a subset of this data set by only looking at the initial parole board interviews. That is, the first time a prisoner appears before the parole board. The final data set has 9580 observations (Not Granted - 6962, Granted - 2618). The variables selected for our model are shown in Table 4.3. Variables were selected based on their perceived relevance to the decision and if a variable had no impact on model performance it was removed. Logistic regression assumption checks were completed. The posterior of each variable was also observed to see if the 95% credible interval contained zero (meaning it has little to no impact on the model).

[2]Data library `https://publicapps.doccs.ny.gov/ParoleBoardCalendar/About?form=datadef#Decision`

Table 4.3: Variable Names and Descriptions

| **Variable Name** | **Variable Description** |
|---|---|
| *Parole Board Decision* | Simplified labels to a binary response: Granted = {Open Date, Granted, Paroled}, Not Granted = {Denied, Not Granted}. |
| *Gender* | Male, Female |
| *Ethnicity* | Black, White, Hispanic, Other |
| *Age* | Years from birth date to interview date. |
| *Crime 1 Class* | Felony codes A, B, C, D and E. A felonies being the most serious and E felonies being the least serious. |
| *Number of Years to Release Date* | Years from interview date to release date. |
| *Number of Years to Parole Date* | Years from interview date to parole eligibility date. |
| *Aggregated Maximum Sentence* | Maximum aggregated amount of time a prisoner must serve for the crimes they are convicted of. |
| *Aggregated Minimum Sentence* | Minimum aggregated amount of time a prisoner must serve for the crimes they are convicted of. |
| *Crime Count* | Number of crimes a prisoner was convicted of under the given sentence (not all criminal history, just crimes for the current prison stay). |
| *Crime 1 Conviction* | Simplified down to the following set: {Possession: Crimes involving possession of an illegal substance or firearm; Grand Larceny: taking of goods in excess of $1000; Assault: Crimes involving assault, excl. sexual assault; DWI: Driving under the influence of drugs or alcohol; Court: Crimes involving court proceedings(e.g., perjury, contempt); Sale: Crimes involving sale of an illegal substance or firearm; Sexual: Any sex related crime (e.g., sexual assault, rape); Fake: Crimes where an individual has faked something (e.g., forgery, identify theft); Death: Any crime where an individual has caused death excl. murder (e.g., manslaughter, homicide); Stalking: including surveillance and harassment; Conspiracy, Murder, Robbery, Arson, Fraud, Kidnapping, Other: All other crimes which do not come under any of the other labels}. Reducing categories in this way is common practice in statistics and is done throughout crime modelling [28]. |

### 4.4.2 Model

We wish to model the Parole Board Decision (response variable) using all other variables as explanatory variables (Table 4.3). Numeric variables are standardised and categorical variables are changed to dummy variables. The model is fitted and posterior distributions are found on a training data set that consists of 80% of the full data set (7664 observations). The performance

measures are assessed for a test data set of observations the model has never seen. The test data set consists of the remaining 20% of the full data set (1916 observations). For a more accurate picture of how the model behaves, we randomly sampled five different testing and training sets and fitted the model separately in each case. We then took the average of the five different accuracy readings produced to get the final values. The structure of the model is shown in Equation 4.6.

$$\begin{aligned} Decision_i &= \beta_0 + \beta_1 \times gender\_male_i + \beta_2 \times age_i + \beta_3 \times num\_years\_release_i \\ &\quad + \beta_4 \times num\_years\_parole_i + \beta_5 \times crime\_count_i + \beta_6 \times agg\_min\_sent_i \\ &\quad + \beta_7 \times agg\_max\_sent_i + \beta_8 \times eth\_hispanic_i + \beta_9 \times eth\_white_i \\ &\quad + \beta_{10} \times eth\_other_i + \beta_{11} \times crime\_class\_B_i + \beta_{12} \times crime\_class\_C_i \\ &\quad + \beta_{13} \times crime\_class\_D_i + \beta_{14} \times crime\_class\_E_i \\ &\quad + \beta_{15} \times crime\_conviction\_assault_i \\ &\quad + \beta_{16} \times crime\_conviction\_burglary_i + \ldots \qquad (4.6) \\ p_i &= \frac{1}{1 + e^{Decision_i}} \end{aligned}$$

All parameters were initialised with a $Normal(0, 0.001)$ prior. All trace plots of the parameters were acceptable.

### 4.4.3 Model Diagnostics

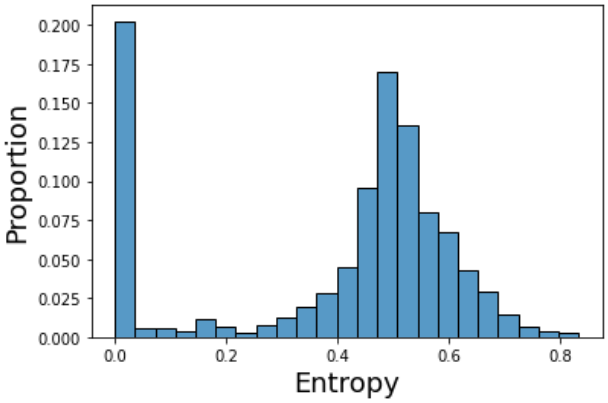
Accuracy readings were taken for the five different test sets and can be found in Table 4.4. The model obtains about 79% classification accuracy overall. The CI accuracy is approximately 84%, with 87% of the CIs being on either side of 0.5, showing that the model is making more certain predictions than predictions that could be either "Granted" or "Not Granted" (corresponding to CI's containing 0.5). The F-score is around 0.87, which is close to one,

showing that the model has relatively good specificity and sensitivity. Figure 5.3a shows the entropy of all observations in a single test set[3]. There are two peaks, one around zero and another around 0.5. From this, we can conclude that our model has some very certain predictions (peak around zero) and some less certain or very uncertain predictions (peak around 0.5). To gain further insight into the behaviour of our model in terms of entropy, Figure 5.3b displays the entropy of correct predictions the model made and Figure 5.3c displays the entropy of incorrect predictions. We can see that for incorrect predictions the large peak at zero is not present (Figure 5.3c), whereas it is still present for correct predictions (Figure 5.3b) meaning our model is less certain with its predictions when it is incorrect. The model looks relatively well-calibrated to the data (Figure 4.4a). The confusion matrix (Figure 4.4b) shows that the model has a high true positive rate, that is, the model is predicting "Not Granted" well, which is to be expected due to the disproportionate amount of "Not Granted" versus "Granted" parole decisions in the data-set. Overall, we believe the model show acceptable behaviour for the proposed task.

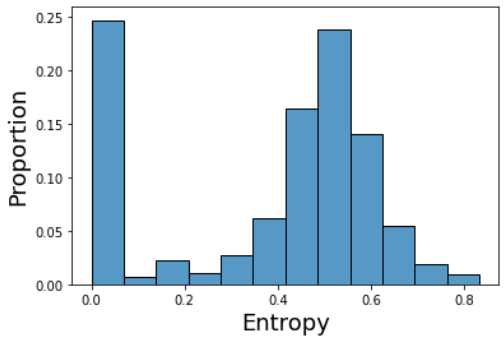
Table 4.4: Average Performance Measures from Five Models.

| **Accuracy Measure** | **Average** |
|---|---|
| *Mean Accuracy* | 79.538% |
| *Mode Accuracy* | 79.498% |
| *Median Accuracy* | 79.51% |
| *AUC Accuracy* | 79.488% |
| *95% CI Accuracy* | 84.542% |
| *Percentage of the 95% CI correct predictions that contain 0.5* | 12.832% |
| *Percentage of the 95% CI correct predictions that are either side of 0.5* | 87.164% |
| *F-Score* | 0.867 |

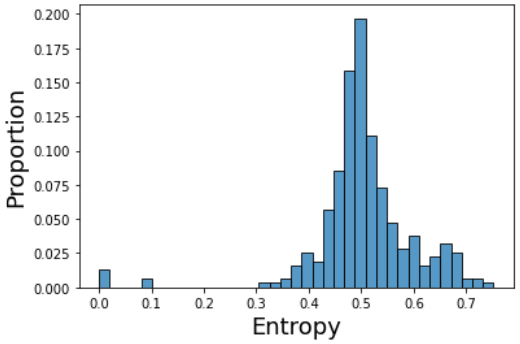
[3]NB:outputs were similar for all five test sets




(a) Histogram of the entropy for all test predictions.




(b) Histogram of the entropy for test predictions where the model made a correct prediction.

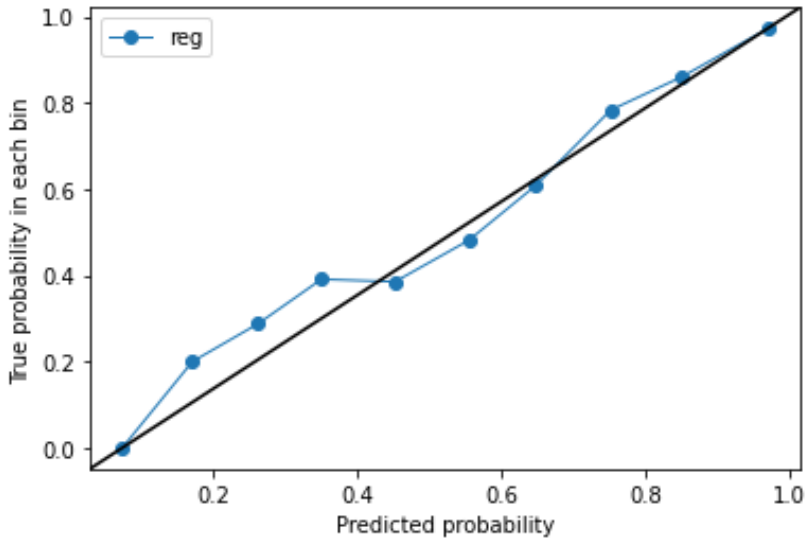



(c) Histogram of the entropy for test predictions where the model made an incorrect prediction.

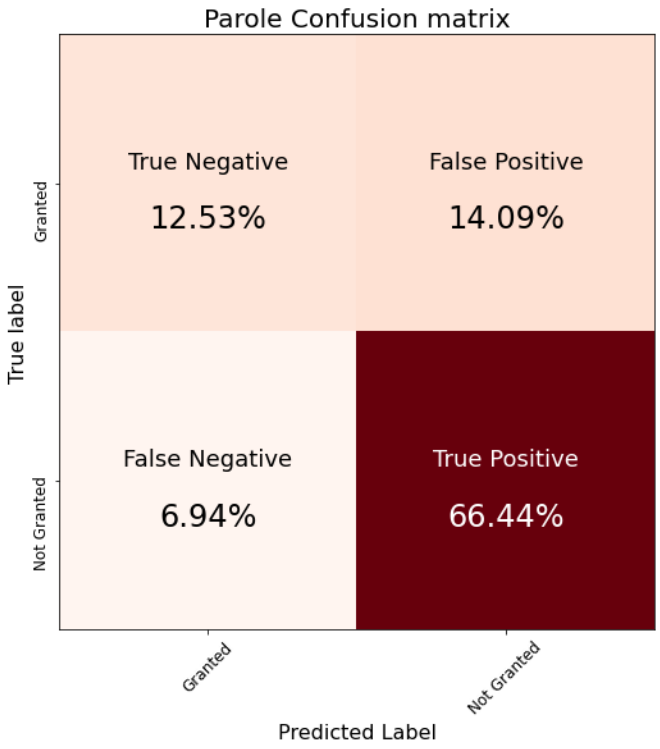
Figure 4.3: Entropy Plots for the Prisoner Re-offending Example

### 4.4.4 Elicited Prior Distribution

After selecting the appropriate model, we can now obtain the elicited prior distribution for a new case. To produce a distribution of expert uncertainty for a single case, we obtain samples of $p_i$, the probability of a prisoner re-committing a crime, using the available information on the prisoner. We do this by sampling 100 times from the posterior distributions of the model parameters. These samples are then used to calculate samples of $p_i$, the probability of a decision $Y_i|X_i$. Then, the method of moments is used to fit a beta distribution to the samples of $p_i$, producing a final distribution capturing uncertainty. An analyst can also choose to fit other distributions to the data by MLE. They can then select the best distribution by the Kolmogorov-Smirnov test [29].




(a) Calibration Plot




(b) Confusion Matrix

Figure 4.4: Model Diagnostic Plots for the Prisoner Re-offending Example

Consider three prisoners: Prisoner 1, Prisoner 2 and Prisoner 3 (the prisoners' attributes are found in Table 4.5). The elicited prior distributions are shown in Figure 4.5. Prisoner 1 yielded a $Beta \sim (74.111, 266.202)$ prior dis-

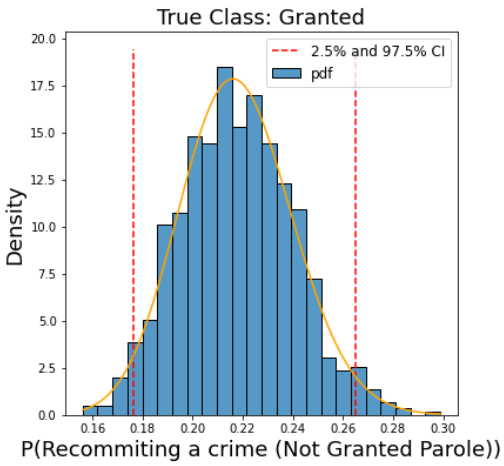
tribution (Figure 4.5a). Prisoner 2 yielded a $Beta \sim (382.491, 154.224)$ prior distribution (Figure 4.5b). Prisoner 3 yielded a $Beta \sim (1181.395, 7.210)$ prior distribution (Figure 4.5c). These elicited distributions can now be used as prior distributions for recidivism for the given individuals and can be used to aid further decision-making.

Table 4.5: Prisoners' Attributes used for the Prisoner Re-offending Example

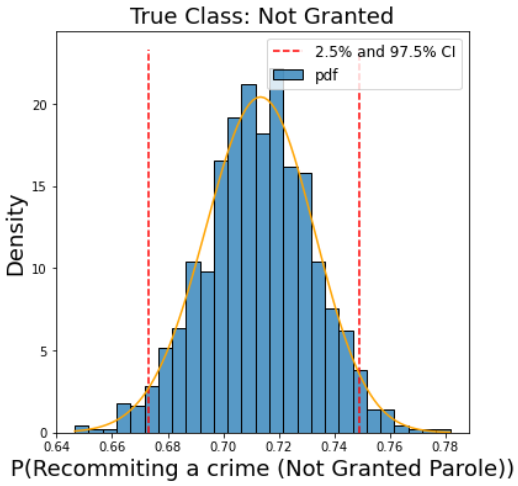
| **Attribute** | **Prisoner 1** | **Prisoner 2** | **Prisoner 3** |
|---|---|---|---|
| Age: | 34 years | 23 Years | 29 years |
| Number of years to release date: | 0 years | 0 years | 1 year |
| Number of years to parole date: | 0 years | 0 years | 0 years |
| Aggregated Maximum Sentence: | 3 years | 3 years | 4 years |
| Aggregated Minimum Sentence: | 1 year | 1 year | 1 years |
| Gender: | Male | Male | Male |
| Ethnicity: | White | Black | White |
| Crime Count: | 1 | 1 | 2 |
| Crime 1 Conviction: | Burglary | Possession | DWI |
| Crime 1 Class: | D | E | E |
| Decision: | Granted | Not Granted | Not Granted |

### 4.4.5 Influential Variables

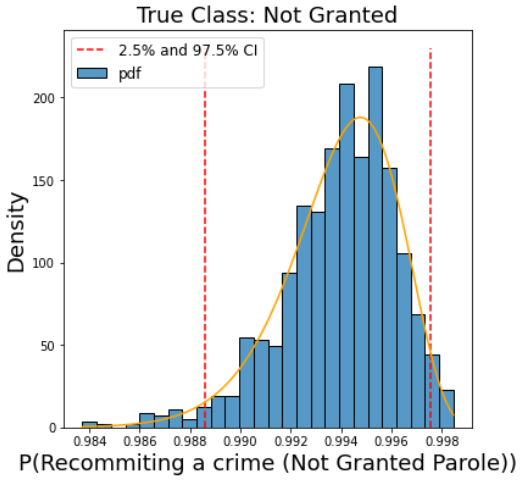
For this example, we have shown how an analyst can elicit a prior distribution from an expert decision-making process using tabular data. However, can an analyst trust that this elicited distribution is reliable? Can they trust the expert's decisions? Could some variables be wrongly influencing decisions? We chose to consider these questions by exploring variables seen in the decision-making process that should not have a cause-effect relationship with the decision. The variables we chose to explore are ethnicity, gender, and age. To explore the effect of these variables, we first created models without these variables and compared them to the original model. Each model was run five times with different testing and training data sets to produce an average of all




(a) Prisoner 1




(b) Prisoner 2




(c) Prisoner 3

Figure 4.5: Prior Distributions for Three Different Prisoners

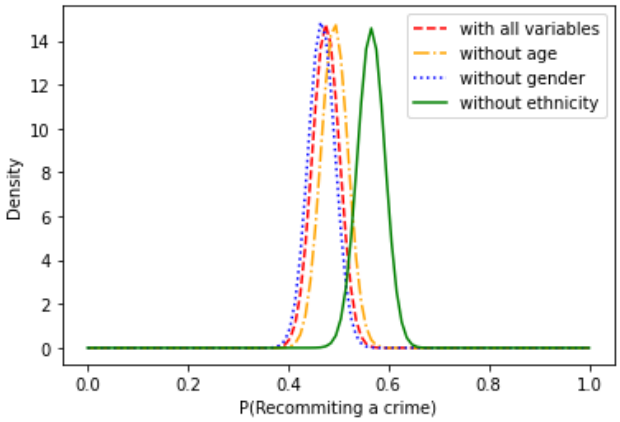
accuracy measures.

The model without ethnicity obtained the lowest average accuracy, and in fact, all five testing data sets gave lower accuracy than the full model (Table 4.6). It is also interesting to see that the model without Ethnicity has a higher percentage of 95% CI correct predictions that contain 0.5. The model without age behaves roughly similar to the full model and the model without gender is only slightly less accurate. We also look at the behaviour of the elicited distribution of a test point from each model (Figure 4.6). It can be seen that for each prisoner the full model and the models without age or gender

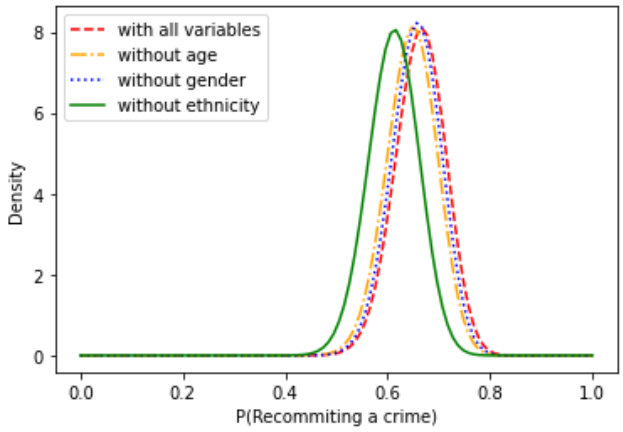
perform similarly, however, the model without ethnicity produces a different distribution (Figure 4.6). This finding is consistent for all prisoners considered. We can further explore the impact of the variable ethnicity by using the full model and looking at a single prisoner and changing their ethnicity (Figure 4.7). Again, there is a clear difference between the different ethnicity's elicited distributions. This shows us that ethnicity has an impact on the decision. Removing ethnicity from the model may reduce bias in the elicited prior distribution, but, it should be noted, that sometimes variables in tabular form can be representing other information that may be valuable to elicit an accurate prior distribution (confounding variables). For example, the variable ethnicity may be a proxy for socioeconomic status [30]. This is a limitation of incomplete tabular data, as an analyst can only assume what this other information is. In this context, it is worth noting that there may be other methods that can go beyond tabular data and allow an analyst to use all the information a decision maker considered to elicit a prior distribution so that all the necessary information is kept in the model to elicit a prior distribution.

Table 4.6: Accuracy Measures of Models where the Variables of Interest are Removed

| **Accuracy Measure** | **Full Model** | **Model without Ethnicity** | **Model without age** | **Model without gender** |
|---|---|---|---|---|
| *Mean Accuracy* | 79.538% | 78.286% | 79.77% | 78.988% |
| *Mode Accuracy* | 79.498% | 78.288% | 79.488% | 78.904% |
| *Median Accuracy* | 79.51% | 78.298% | 79.72% | 78.988% |
| *AUC Accuracy* | 79.488% | 78.298% | 79.72% | 78.978% |
| *95% CI Accuracy* | 84.542% | 85.564% | 84.394% | 84.3% |
| *Percentage of the 95% CI correct predictions that contain 0.5* | 12.832% | 17.608% | 12.074% | 14.168% |
| *Percentage of the 95% CI correct predictions that are either side of 0.5* | 87.164% | 82.388% | 87.904% | 85.83% |
| *F-Score* | 0.867 | 0.857 | 0.868 | 0.86356 |




(a) Prisoner 4

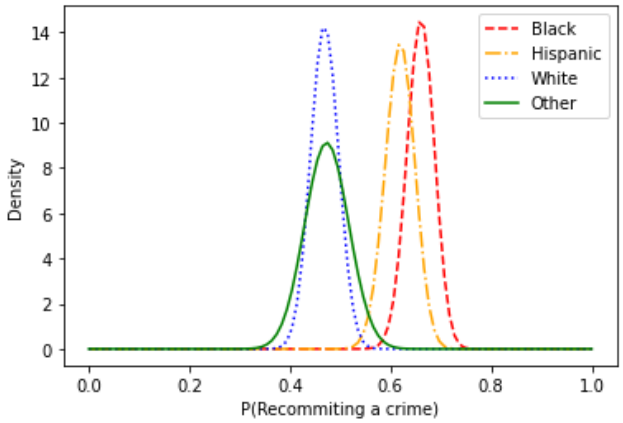



(b) Prisoner 5

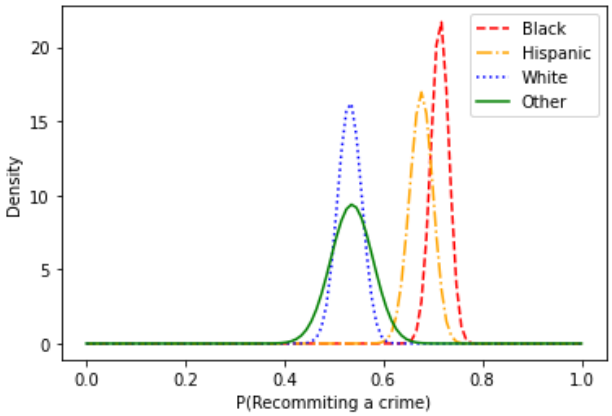
Figure 4.6: Elicited Distributions for the Four Different Models for Different Prisoners.

### 4.4.6 Summary

This example shows how to elicit expert uncertainty present when considering whether a prisoner will re-commit a crime upon release, using a Bayesian logistic regression model to model parole board decision-making. The proposed process enables an analyst to also observe the impact of variables that may be influencing the decisions. The example has limitations; the parole board usually makes its decisions based on a report submitted by a prisoner's case worker. The only available data considered in the example was tabular data, which does not provide all the information that would be in the report. It would be interesting to see if modelling the report data would provide dif-




(a) Prisoner 4




(b) Prisoner 5

Figure 4.7: Elicited Distributions for The Same Test Point but with Ethnicity Variable Changed

ferent results to those obtained above. Also, as with any elicitation method, there may be questions regarding the accuracy of the elicited prior distributions. For this example, we assume that all historic decisions are indicative of future decisions. This may not always be appropriate and may create inaccuracies. The accuracy of elicited prior distributions is an ongoing concern of the prior elicitation field [31] and should be a continual path for future research (discussed further in Section 5.5).

## 4.5 Conclusions and Future Work

We introduce a new method to elicit prior distributions for an event, by modelling an expert decision-making task. We assume that a decision, $Y$, is closely related to the event $A$ so that samples from $P(Y|X, \theta)$, for different values of $\theta$, can be used to approximate the prior distribution for $A$ given a particular case $X$. This method allows an analyst to elicit a prior distribution from a real-world expert decision-making process, without the expert needing knowledge of probability concepts. This method can also be easily implemented for multiple experts where a decision is made in consensus because it models one decision, no matter if an individual or group makes the decision. We introduced this method with an example of recidivism using tabular data. This example used Bayesian logistic regression to model the parole board decision-making process. Once an appropriate model was fitted, samples from the posterior distributions of the parameters were taken to form a distribution that can be used as a prior distribution for recidivism.

It is important to note that our approach is valid beyond just regression models. As summarised in Section 5.3, we use predictive modelling of a decision-making task to infer the expert prior. Therefore, as long as i) there exists an event $A$, a decision $Y$ (that reflects the decision maker's beliefs about the value $A$ will take) and related information $X$ to make the decision $Y$, and ii) there is a sufficient number of recorded past decisions to infer a posterior distribution on $\theta$, this prior elicitation approach can work. There is also an implied assumption that the past decision-making is still relevant.

We introduced this method with a simple example that only requires binary outcomes. However, there are many situations where the decisions are not binary but categorical. In such situations, an analyst can use Bayesian multinomial logistic regression ([32, 33]) to elicit distributions. This can be done in a similar way to that discussed in Section 5.3 by producing a sample of $P(Y = ResponseA)$ by sampling from the posterior distributions of the fitted

model parameters. However, here an analyst must truly understand which of the decisions aligns with the event uncertainty that they wish to elicit. This process may require a lot of careful consideration and if not done correctly may create an inaccurate distribution. If multiple decisions could align with the desired uncertainty an analyst wishes to elicit, then an analyst could create a new binary response variable, $Z$. Where $Z = 1$, includes all decisions that align with the desired uncertainty (that is if decision $D = \{Response1, Response2\}$ and $Z = 0$, includes all the decisions that do not align with the desired uncertainty (that is, $D = \{Response3, Response4\}$, allowing an analyst to use a simple Bayesian logistic model.

Using this method also enables an analyst to explore variables that may be strongly influencing the decision-making process. What to do with this information should be a topic of future research. Should an analyst remove this information, or should it be shared with the experts to help train for future decision-making? A limitation of the logistic regression example considered in this paper is that the use of tabular data makes it challenging for an analyst to truly ascertain what is influencing the decision-making as this type of data only provides limited information and is often not what an expert would use to make their decisions. It would be interesting to explore modelling decision-making tasks that involve more complex data, such as images or reports. Basic statistical models cannot perform these tasks; instead, machine-learning approaches will have to be implemented. It would also be intriguing to investigate how this approach can be expanded for scenarios where time may impact the decision-making process and how time affects the elicited distributions. We acknowledge that this method will not work where there is no data (or not enough data) from an appropriate decision-making task. A concern in the field of prior elicitation is how accurate the elicited prior distribution is in terms of the true prior for an event; further research could be taken to see how accurate this method of prior elicitation is and, if there is a method to calibrate the

elicited distribution against any biases introduced by the experts (See example in [31]). By using a number of accuracy measures (discussed in Section 4.3) and the full machinery of Bayesian inference to model past decision-making, we know how well the inferred priors capture the expert's uncertainty in a manner that is consistent with their past decision-making.

It's important to note that using all past decisions as a predictor of future decisions may not always be appropriate and could lead to inaccuracies. An analyst could select only decisions that would be considered relevant. However, if an analyst does include all past decisions to infer a prior distribution, then calibration techniques could be utilised. If there exist cases where the outcome of the event $A$ has been observed, these could potentially be used to calibrate the elicited prior distribution.

Overall, although we hope to have argued successfully that the proposed method is a promising candidate for prior elicitation in practical applications, further research should be performed to improve the practicality and generality of the approach.

# References

[1] Michael J Zyphur and Frederick L Oswald. Bayesian estimation and inference: A user's guide. *Journal of Management*, 41(2):390–420, 2015.

[2] Julia R Falconer, Eibe Frank, Devon LL Polaschek, and Chaitanya Joshi. Methods for eliciting informative prior distributions: A critical review. *Decision Analysis*, 19(3):189–204, 2022.

[3] Anthony O'Hagan, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. Uncertain judgements: eliciting experts' probabilities. 2006.

[4] Lionel A Galway. Subjective probability distribution elicitation in cost risk analysis: A review. 2007.

[5] David Jenkinson. The elicitation of probabilities: A review of the statistical literature. 2005.

[6] Owen Thomas, Henri Pesonen, and Jukka Corander. Probabilistic elicitation of expert knowledge through assessment of computer simulations. *arXiv preprint arXiv:2002.10902*, 2020.

[7] Christopher J Casement and David J Kahle. Graphical prior elicitation in univariate models. *Communications in Statistics-Simulation and Computation*, 47(10):2906–2924, 2018.

[8] Joseph Kadane and Lara J Wolfson. Experiences in elicitation: [read before the royal statistical society at a meeting on'elicitation 'on wednesday, april 16th, 1997, the president, professor afm smith in the chair]. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):3–19, 1998.

[9] Chen Wang and Vicki M Bier. Expert elicitation of adversary preferences using ordinal judgments. *Operations Research*, 61(2):372–385, 2013.

[10] Robert L Winkler. The quantification of judgment: Some methodological suggestions. *Journal of the American Statistical Association*, 62(320):1105–1120, 1967.

[11] Robert T Eckenrode. Weighting multiple criteria. *Management science*, 12(3):180–192, 1965.

[12] Ward Edwards and F Hutton Barron. Smarts and smarter: Improved simple methods for multiattribute utility measurement. *Organizational behavior and human decision processes*, 60(3):306–325, 1994.

[13] Daniel Kahneman, Stewart Paul Slovic, Paul Slovic, and Amos Tversky.

*Judgment under uncertainty: Heuristics and biases*. Cambridge university press, 1982.

[14] Anthony O'Hagan. Expert knowledge elicitation: subjective but scientific. *The American Statistician*, 73(sup1):69–81, 2019.

[15] Irving Lester Janis. *Groupthink*. Houghton Mifflin Boston, 1983.

[16] S James Press. *Subjective and objective Bayesian statistics: principles, models, and applications*. John Wiley & Sons, 2009.

[17] Lorenzo Belenguer. Ai bias: Exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. *AI and Ethics*, 2(4):771–787, 2022.

[18] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2016.

[19] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications, 2018.

[20] Paul A Jargowsky. Omitted variable bias. *Encyclopedia of social measurement*, 2:919–924, 2005.

[21] Nikolaj Tollenaar and Peter GM van der Heijden. Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(2):565–584, 2013.

[22] Jonathan Caulkins, Jacqueline Cohen, Wilpen Gorr, and Jifa Wei. Predicting criminal recidivism: A comparison of neural network models with statistical methods. *Journal of Criminal Justice*, 24(3):227–240, 1996.

[23] Rolando de la Cruz, Oslando Padilla, Mauricio A Valle, and Gonzalo A Ruz. Modeling recidivism through Bayesian regression models and deep neural networks. *Mathematics*, 9(6):639, 2021.

[24] Peter Schmidt and Ann Dryden Witte. Predicting criminal recidivism using 'split population'survival time models. *Journal of Econometrics*, 40(1):141–159, 1989.

[25] Yutaka Sasaki et al. The truth of the f-measure. *Teach tutor mater*, 1(5):1–5, 2007.

[26] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[27] David JC MacKay, David JC Mac Kay, et al. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[28] Australian Bureau of Statistics. Australian and New Zealand Standard Offence Classification (ANZSOC). `https://www.abs.gov.au/statistics/classifications/australian-and-new-zealand-standard-offence-classification-anzsoc/2011`, 2011. [Online; accessed 21-February-2023].

[29] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[30] American Psychological Association et al. Ethnic and racial minorities & socioeconomic status. *American Psychological Association. http://www.apa. org/pi/ses/resources/publications/factsheet-erm. aspx [accessed October 28, 2011]*, 2017.

[31] Tommi Perälä, Jarno Vanhatalo, Anna Chrysafi, et al. Calibrating expert assessments using hierarchical gaussian process models. *Bayesian analysis*, 15(4):1251–1280, 2020.

[32] Jared D Fisher and Kyle R McEvoy. Bayesian multinomial logistic regression for numerous categories. *arXiv preprint arXiv:2208.14537*, 2022.

[33] Sean M O'brien and David B Dunson. Bayesian multivariate logistic regression. *Biometrics*, 60(3):739–746, 2004.

# Chapter 5

# Utilising Deep Learning to Elicit Expert Uncertainty

Recent work [1] has introduced a method for prior elicitation that utilizes records of expert decisions to infer a prior distribution. While this method provides a promising approach to eliciting expert uncertainty, it has only been demonstrated using tabular data, which may not entirely represent the information used by experts to make decisions. In this paper, we demonstrate how analysts can adopt a deep learning approach to utilize the method proposed in [1] with the actual information experts use. We provide an overview of deep learning models that can effectively model expert decision-making to elicit distributions that capture expert uncertainty and present an example examining the risk of colon cancer to show in detail how these models can be used.

## 5.1 Introduction

When choosing from a set of actions, it is important to understand the level of uncertainty surrounding an event in order to make informed decisions and assess risks [2]. Analyzing numerical data or visual representations of uncertainty can assist decision-makers and risk assessors in providing insightful evaluations. However, quantifying uncertainty in this way can be a difficult task, especially where there is limited data on the event. This is where expert

knowledge elicitation is needed, defined here in the statistical sense where the goal is to elicit a probability distribution that encapsulates expert uncertainty regarding some uncertain quantity or event, $E$ [3, 4].

Knowledge elicitation techniques can be technical and time-consuming to apply. Standard techniques of knowledge elicitation focus on interviewing experts to obtain probability distributions reflecting their uncertainty. An analyst can interview experts and ask them questions on the probability distributions (Direct Interrogation) [3, 5, 6] to capture their uncertainty. Questions may be on the probabilities [3], such as, "What is the probability that the proportion of $E$ is less than or equal to 0.8?"(i.e., $P(E \leq 0.8)$); or focus on the quantiles of the distribution [7], "at what value is the probability of $E$ equally likely to be less than or greater than that value?" (i.e., estimate the median)[8]. Direct Interrogation methods may induce bias from the types of questions asked (e.g. Anchoring and Adjusting, where giving a value in the question may bias the expert's response [9]) and require statistical expertise on probabilities and distributions, which is often not feasible. To reduce the statistical knowledge required from experts, an analyst may alternatively ask them to perform hypothetical decision-making tasks (Indirect Interrogation) [10, 11, 12, 13, 8]. Some examples are getting experts to rank the likelihood of events or getting them to place bets on which events they think are more likely. Hypothetical decision-making tasks like these often do not place significant importance on the accuracy of the decisions, so they can potentially produce inaccuracies in the elicited distributions [8].

As highlighted by [3], not only can expert knowledge elicitation be used for risk assessments and decision-making, it can also be profitably applied in two statistical contexts: experimental design and elicitation of informative prior distributions for Bayesian inference. In the context of Bayesian statistics, [1] introduces a method for prior elicitation that combats some of the

challenges faced in knowledge elicitation. Their method focuses on eliciting informative priors from expert decision-making tasks that are repetitive and carried out regularly, often under real-life circumstances. They explore the relationship between a rare event $E$ and a decision, $Y$, made by experts, selecting a decision-making task by considering whether the decision $Y$ reflects the uncertainty in the event $E$. [1] illustrates this method with the event of an individual prisoner recommitting a crime upon release from prison, where there is limited data on this event occurring, and this event is hoped never to happen. A decision-making process that reflects the uncertainty around said event is the parole board decision-making process. The parole board assesses the risk of a prisoner committing a crime upon release when making their decision. Falconer et al. explain how to model the decision-making process using Bayesian inference to elicit a probability distribution capturing the uncertainty of the decision makers. Explained simply, if $Y \sim Bernoulli(p)$ and there is tabular data available from the decision-making task, then a Bayesian logistic regression model can be used to model $Y$, with $p = f(X, \beta)$, where $X$ is the information describing an event $E$. An analyst can use this model to elicit a distribution for $p$ by sampling multiple times from the posterior distribution of model parameters, $\beta$, to obtain a sample of $p$. The method of moments can be used to fit a distribution for $p$. The prior probability distribution for an event $E$ is obtained by inputting the information for the event, $X^*$, into the logistic regression model, sampling repeatedly, and using the distribution found using the method of moments as the prior probability distribution for $E$. This method eliminates the statistical knowledge required to elicit a prior from an expert, as the expert is just completing their usual decision making. Ideally, all decisions are made in real-life circumstances, meaning more thought is put into making the right decision.

Although initially introduced in a Bayesian context, this method has wider applications in the field of knowledge elicitation. By relying solely on the data

used for decision making instead of requiring interaction with experts, we can simplify the process of eliciting expert uncertainty. This approach is particularly useful when experts are already engaged in daily decision-making tasks. However, although promising, this method has only been investigated in the context of tabular data so far [1]. Tabular data can introduce bias by only considering measurable/recorded characteristics [14]; this means that information that is important to the decision-making process may be lost. Real-life decisions are usually made on more complex data, such as images or reports. For example, following the example in [1], the parole board considers whether a prisoner will commit a crime upon release from prison to decide whether or not to grant parole. To come up with this decision, the board reviews a report created by the prisoner's case worker and testimonials from others. This report contains all the information the board needs to make an informed decision. To obtain the appropriate distribution, we must first model the decision-making process. To do this, we must use all the information available to the decision-maker to elicit their uncertainty. Obtaining a distribution from complex data, like a report, cannot be achieved by standard statistical models. Instead, machine learning and artificial intelligence models can be utilized. In particular, deep learning is a branch of machine learning that allows analysts to model complex data. Although deep learning has been used for both uncertainty quantification [15] and for modelling decision-making [16], it has never been used, as far as we are aware, to elicit expert knowledge in the form of a prior probability distribution.

In this paper, we extend the work in [1] using deep learning to obtain a model that captures an expert's uncertainty in the form of a probability distribution. Before delving into the specifics, it is essential to understand the fundamental concepts of uncertainty, which will be discussed in Section 5.2. Section 5.3 outlines how to capture uncertainty using deep learning methods. We will then outline an example in Section 5.4 that assesses whether or not

a patient is at risk of developing cancer by modelling a histopathology data set [17] with the majority diagnoses from seven pathologists. This example showcases the power of deep learning models to model complex data to obtain a probability distribution.

## 5.2 Uncertainty

Understanding expert uncertainty can be a challenging task. When it comes to elicitation of uncertainty, an analyst must first identify the type of uncertainty they wish to elicit and ensure that their models reflect that uncertainty. In particular, considering the method proposed in [1], it is important for the analyst to balance two factors: accurately mimicking decision-making and ensuring that the elicited distributions align with the expected behaviour with respect to the types of uncertainty deemed relevant. There are two primary types of uncertainty to consider: aleatoric uncertainty and epistemic uncertainty [18, 19]. Aleatoric uncertainty, also referred to as objective uncertainty, arises from random variation. An example would be tossing a fair coin. At the time of the coin toss, we can never be 100% certain that the coin will land on heads due to the random nature of the event. Aleatoric uncertainty is irreducible, meaning it cannot be eliminated from the system. In contrast, epistemic uncertainty, also referred to as subjective uncertainty, is reducible and stems from the lack of knowledge about an event. Epistemic uncertainty occurs commonly when we have limited information on an event and is decreased when more information is provided. Although identifying what type of uncertainty is present in a particular task may seem straightforward, it is not. Identifying uncertainty is dependent on the task at hand. What is labeled as a particular type in one study may be labeled as the other type in a different study [20]. Kiureghian and Ditlevsen [19] suggest that classifying uncertainty type may be more based on the immediate situation, which uncertainties can

be immediately reduced and which may be more difficult to reduce in the near future. The reader should refer to [19, 20] for more information.

The work in [1] aimed to quantify the aleatoric uncertainty of a released prisoner committing another offense by creating a model of the parole board decision-making process, which considers the uncertainty of a prisoner committing a crime after being released. To fit an appropriate model, this work explored model accuracy measures that take into account the elicited uncertainty and ensure that the model does not elicit certain distributions where it should be uncertain. The primary concern in the parole board application is to quantify aleatoric uncertainty based on the history of the prisoner. By considering parole board decisions, it may be impossible to completely eliminate epistemic uncertainty even if complete historical data on the prisoner is available because of expert bias (which is shown in the analysis performed in [1]) and the difficulty of representing data comprehensively in tabular form. The goal should be to reduce the epistemic uncertainty as much as possible to highlight the aleatoric uncertainty. This may be achieved by using bias adjusting techniques or using other data types [1, 21, 22, 23].

In certain scenarios, eliciting expert epistemic uncertainty is a priority [24, 25]. A prime example is the uncertainty surrounding a patient's diagnosis. One expert may provide a diagnosis that differs from that of another expert due to differing knowledge, experience, and training or insufficient time to form a comprehensive opinion. While obtaining the opinion of a single expert can be useful, acquiring the uncertainty of a group of experts adds crucial insight and context. If we can gather the group's uncertainty, we can better understand the uncertainty of a particular diagnosis based on the group of experts consulted. The importance of obtaining this uncertainty cannot be overstated, especially when it comes to making informed decisions about a patient's future. In Section 5.4, an example is presented that showcases the use

of eliciting epistemic uncertainty. The example involves seven experts who are diagnosing patients, but they do not always agree on a single diagnosis. We can use the expert agreement levels to check if our model is behaving appropriately in eliciting epistemic uncertainty distributions (Section 5.4). Where we have opposing opinions, we expect our model to produce distributions that exhibit less certainty than those for patients who had a full agreement by experts.

Our objective in this paper is to use deep learning to elicit the experts' uncertainty, taking a conservative approach when deciding whether the model is eliciting appropriate distributions. The goal should be for the model not to produce a narrow distribution where the probability mass is centred near zero or one (the model is 'certain' in its outcome) in cases where the experts exhibit uncertainty about the outcome. If the opposite occurs, i.e. the model produces a distribution with a wide credible interval possibly containing 0.5 (the model is 'uncertain') when all experts agree, this is less concerning and consistent with a conservative modelling approach. We would rather the model take this conservative approach than be overly certain and not follow the behaviour of the experts in cases where they are uncertain.

In Section 5.4.3, we outline measures that we use to ensure our model behaves appropriately for the specified task.

## 5.3 Deep Learning

Understanding the human decision-making process is of interest in the field of Artificial Intelligence (AI) [26]. Deep learning is a sub-field of AI that utilizes deep neural networks to learn. A neural network (NN) structure is loosely based on the human brain and how it processes information [27]. It has layers of neurons that process information and pass it to the next layer until it reaches an output layer. This basic NN structure can be built upon to make more complex model architectures with many layers that can process

more complex data, such as images and reports [28, 29]. The most common deep learning models are deterministic, but to quantify uncertainty, we need to make the deep learning models probabilistic.

### 5.3.1 Probabilistic Deep Learning

The key to obtaining a representation of uncertainty from a deep learning model is to apply some probabilistic features to the models. These probabilistic features are commonly used in deep learning to obtain the uncertainty surrounding model predictions by considering the uncertainty in the model parameters given the finite amount of training data available [30]. Two main approaches can be applied to deep learning models to elicit the corresponding probability distributions: Bayesian neural networks (BNN) and Monte Carlo (MC) Dropout.

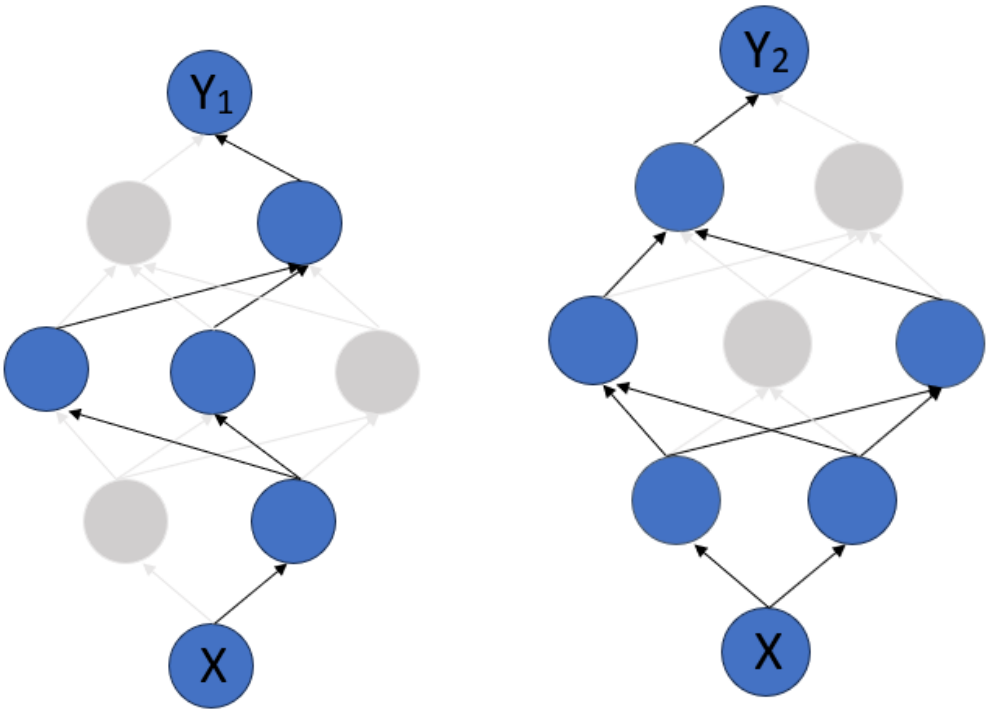
BNNs place priors on the parameters of the neural network and learn the posterior distributions of these parameters [31]. Classical Markov Chain Monte Carlo (MCMC) methods to obtain samples from the posteriors of the parameters can be used in Bayesian deep learning, but they are computationally expensive (require large memory and a lot of time) [32] and are often challenging to implement. Applying MCMC methods to large data sets for complex networks is often impossible, given current technology. Variational inference is more commonly used in BNNs, as it is easier to implement and does not require as much memory and time as MCMC methods [32]. Variational inference is a method to approximate the posterior distribution, $P$, by taking a distribution, $Q$, from a family of distributions of a simpler form than $P$ [33]. The goal is to find a $Q$ that minimises the Kullback-Leibler divergence (Equation 5.1). When applied to a neural network, variational inference finds the parameters, $\theta$ of the approximate distribution on the weights, $q(w|\theta)$ [34]. Bayes by backdrop is a method that can be used to train a network that applies variational inference [34]. MCMC methods and variational inference both have their spe-

cific applications in neural networks, [32] give advice on which method to use for a given deep learning task. They state: "Thus, variational inference is suited to large data sets and scenarios where we want to quickly explore many models; MCMC is suited to smaller data sets and scenarios where we happily pay a heavier computational cost for more precise samples. For example, we might use MCMC in a setting where we spent 20 years collecting a small but expensive data set, where we are confident that our model is appropriate, and where we require precise inferences. We might use variational inference when fitting a probabilistic model of text to one billion text documents and where the inferences will be used to serve search results to a large population of users".

$$D_{KL}(Q||P) = \sum Qlog(\frac{Q}{P}) \tag{5.1}$$

MC-Dropout is a simple method that can produce the desired distribution. It applies random dropout to the layers of the neural network. Dropout is a function that randomly eliminates nodes from the neural network forward calculation; each forward pass will eliminate different nodes. Nodes are eliminated with a probability, $q_i$, specified as a hyperparameter by the user when the model is built. Although initially not described as a probabilistic method, MC-Dropout, in fact, provides a Bayesian approximation [35]. Gal and Ghahramani show that dropout mathematically approximates a probabilistic Gaussian process by minimizing the Kullback-Leibler divergence between an approximate distribution and the posterior of a deep Gaussian process [35]. An analyst can obtain a distribution by applying dropout not only at training but also when making predictions, running the input through the model multiple times, with each run randomly selecting different nodes to drop out (Figure 5.1). [35] not only show that dropout is a Bayesian approximation but, crucially, also improves model performance compared to variational inference.

The key step to obtain a probability distribution from decision-making




(a) First pass through a NN with dropout layers.

(b) Second pass through a NN with dropout layers.

Figure 5.1: Multiple passes through a NN with dropout layers for a single input will produce different results.

tasks through deep learning is to furnish the neural network with a mechanism to produce an output that can be viewed as a probability ( i.e., a value between zero and one). This can be achieved by applying a softmax function, $\sigma$, (Eq. 5.2) to the outputs of the NN, $z_i$ (or a sigmoid function for binary classification). If the model outputs a probability, $p_i$, of the decision, $Y_i$, we can obtain a sample of $p_i$ through the probabilistic deep learning methods described above. Subsequently, the Method of Moments can be used, as in [1], to fit a Beta density function to the samples of $p_i$, producing a final distribution capturing uncertainty.

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}, \tag{5.2}$$

To produce a probability distribution, analysts have various deep learning

methods at their disposal. We recommend starting with a deep learning model that includes dropout layers, as it is a simpler and less computationally expensive approach that still delivers good model performance compared to other methods [35]. It is important to note that these deep learning methods are complex models with numerous parameters; analysts must carefully consider model accuracy measures when selecting the best method while taking into account the time and memory consumed by each model. While there is still much research to be done in making these methods compatible with cutting-edge programming tools, the field of deep learning is constantly evolving, and better techniques are constantly becoming available.

## 5.4 Diagnosis Example

To make the proposed mechanism for eliciting experts' uncertainty using deep learning more concrete, we will run through a simple example with the reader to show the practical uses for prior elicitation.

### 5.4.1 Data

Let $A$ be the event that an individual develops colon cancer in the future. We wish to elicit a probability distribution reflecting expert uncertainty regarding the probability of this event occurring, given imagery from histopathology. To elicit this distribution from specialists, we can model the data from the histopathology decision-making process [17]. The data set contains 3,152 images of microscope slides of colon tissue (hematoxylin and eosin (H&E)-stained Formalin-Fixed Paraffin-Embedded (FFPE) of colorectal polyps) from the Department of Pathology and Laboratory Medicine at Dartmouth-Hitchcock Medical Center (DHMC) [17]. Each image is assigned a diagnosis of either Hyperplastic Polyp (HP) or Sessile Serrated Adenoma (SSA). The SSA diagnosis is the presence of a pre-cancerous cell which can turn into cancer if untreated [36, 17]. Each image was observed by seven different pathologists,

and the majority vote was taken as an image's diagnosis. The data set also contains the number of pathologists who agreed with the diagnosis, showing that the diagnosis of SSA is not a certain decision and there is uncertainty among specialists in the decision. Pathologist agreement level is not included in the model building (only images and diagnosis are used) but can be used to assess model performance. There are 990 images labelled as SSA in the data set ($\approx 31\%$ of the full data set).

This data, and similar histopathology data sets, have been modelled by deep learning models in the past under the context of image classification [17, 37]. The goal of these classification models is for the model to correctly identify the presence of SSA in every image. We use this data set in the context of knowledge elicitation, where the goal is to accurately quantify expert uncertainty. To do this, we can follow the basic structure of one of the models used in previous research [17] and expand this model with the probabilistic features discussed in Section 5.3 to obtain a distribution that captures uncertainty. After training this model, we then, in Section 5.4.3, complete a set of model diagnostics to make sure our model is aligned with the goal of eliciting expert uncertainty.

### 5.4.2 Model

[17] models this data with a Resnet18 model. Resnet18 is a residual learning model that utilizes 18 convolutional layers, suitable for processing imagery. Residual learning solves the problem of vanishing gradients that can be found in large multi-layer networks, providing a method that will quickly train these networks and obtain better accuracy [38]. Residual models contain "blocks" of small neural networks that learn features and utilise "shortcut connections" [39]to perform identity mapping. That is, where the outputs of the previous block are added to the output of the current block[38]. For more information on residual learning refer to [38]. To fulfil the required probabilistic feature

to output distributions, dropout layers were added between each block of the model and a sigmoid function was applied to the output layer (See Figure 5.2). Each image was scaled down to 100 x 100, and pixel values were normalized. The initial learning rate was set to $10^{-3}$; after the first ten epochs, this was reduced by 0.01 for each epoch. The network was trained for 100 epochs. The batch size for the model was set to 32, and stochastic gradient descent was used as the optimizer. The model was trained with the Binary Cross Entropy (BCE) loss function in the Pytorch library [1].

### 5.4.3 Model Performance

We use the performance measures outlined in [1] (Table 5.1) to assess whether the model is appropriate for the task of uncertainty elicitation. To improve the reliability of the estimates of performance, the model was run ten times with different training and test sets. Accuracy readings were taken for the ten different test sets, and the average of the ten results can be found in Table 5.2.

The model obtains roughly 78 % model accuracy over mean, median, mode and AUC performance measures. For the 95 % credible interval accuracy, we get an accuracy measure of 92.57%; roughly 60 % of these credible intervals are on either side of 0.5, with the other 40 % containing 0.5. This shows us that our model is making some certain and some less certain decisions. We can also observe the behaviour of the entropy of test points in the testing data set (Figure 5.3). Entropy is a measure of uncertainty, which can give us more information on how well the model captures uncertainty. Values closer to zero indicate a narrower distribution, and values closer to one indicate a wider distribution [42]. We obtained a histogram of entropy values for all test points, a histogram of entropy values for test points that assigned the correct label by the 95% credible interval and a histogram of entropy values that

[1] All code for model training and testing can be found at `https://github.com/jrg2223/Utilising-Deep-Learning-to-Elicit-Expert-Uncertainty`

Table 5.1: Model Diagnostics Descriptions Taken From [1]

.

| **Name** | **Description** |
|---|---|
| *Mean Accuracy* | Percentage of correct predictions the model makes by using the mean of the sampled probabilities $p_i$ for each observation. |
| *Mode Accuracy* | Percentage of correct predictions the model makes by using the mode of the sampled probabilities $p_i$ for each observation. |
| *Median Accuracy* | Percentage of correct predictions the model makes by using the median of the sampled probabilities $p_i$ for each observation. |
| *Area Under Curve (AUC) Accuracy* | Percentage of correct predictions the model makes by taking the largest area either side of 0.5 as the measure to form the model prediction. |
| *95% Credible Interval (CI) Accuracy* | Percentage of correct predictions the model makes by observing the 95% CI of $p_i$. If the 95% CI contains 0.5 then the assigned label can be either "Accept" or "Reject" and is a correct prediction. If the 95% CI is contained below 0.5 and the true label is "Accept" then it is a correct prediction. If the 95% CI is contained above 0.5 and the true label is "Reject" then it is a correct prediction. |
| *Percentage of the 95% CI correct predictions that contain 0.5.* | This will allow the analysts to see how many central distributions are elicited. |
| *Percentage of the 95% CI correct predictions that are either side of 0.5.* | This will allow the analysts to see how many skewed distributions are elicited. |
| *F-Score [40]* | A measure which shows the specificity (true negative rate) and sensitivity (true positive rate) of the model. The mean of the samples of $p_i$ is used to assign labels. The highest possible value of an F-score is 1.0, indicating perfect specificity and sensitivity, and the lowest possible value is 0, if either the specificity or the sensitivity is zero. $$F = 2\frac{specificity \times sensitivity}{specificity + sensitivity}$$ . |
| *Confusion Matrix [41]* | Shows the percentage of the mean predictions by whether the prediction is a true negative, true positive, false negative or false positive, showing the specificity and sensitivity of the model. The mean of the samples of $p_i$ is used to assign labels. |
| *Entropy [42]* | A measure of the amount of uncertainty in a distribution. A narrow distribution will give a value close to zero, and a wide distribution will give a value closer to 1. To make sure the model is behaving correctly, it will be helpful to observe a histogram of all entropy values for the training set, as well as observe the histograms of the entropy values of correct and incorrect predictions separately. |
| *Calibration Plot* | A calibration plot shows how well the prediction probabilities match the true percentage probabilities of the data. The mean of the samples of $p_i$ is used as prediction probabilities. |

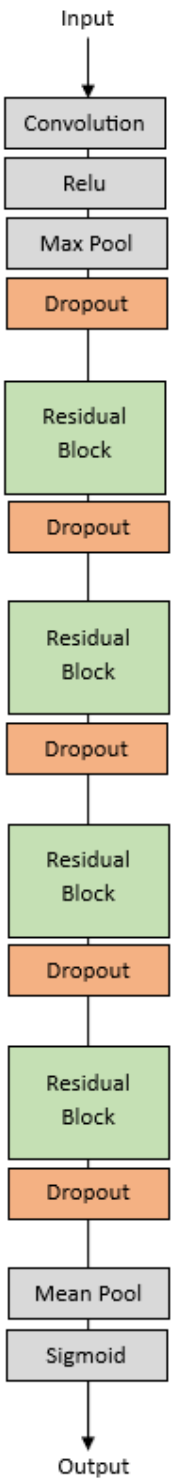


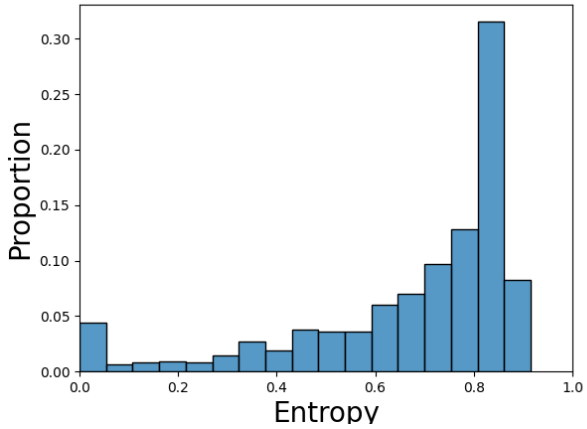

Figure 5.2: Basic Model Structure and Incoporation of Dropout Layers

assigned the incorrect label by the 95% credible interval. Figure 5.3a shows clearly that the model is making some certain (values close to zero) and some uncertain (values close to one) predictions. There is a peak at zero and then an exponential increase in values up to around 0.9. The histogram of correctly

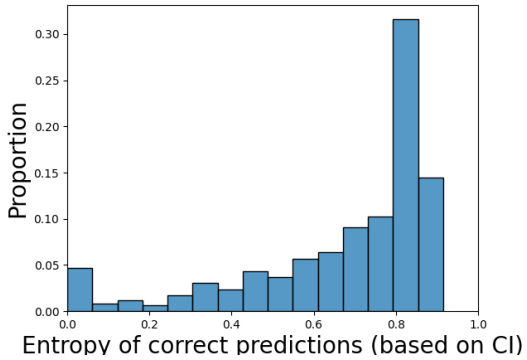
labeled test points shows roughly the same distribution of entropy values as the plot of all test points (Figure 5.3b). The histogram of incorrectly labeled test points no longer has a peak at zero. Instead, a small peak at 0.4, and most points lie between 0.6 and 0.8 (Figure 5.3c. This shows that our model is not certain when it assigns an incorrect label, placing a wider distribution on these points. Our model is well calibrated to the data (Figure 5.4b) and has reasonable sensitivity and specificity (Figure 5.4a, and F-Score in Table 5.2), when using the mean of our probability samples to assign a prediction (Table 5.2). This model behaves reasonably well for the task of expert knowledge elicitation.

Table 5.2: Average Model Performance Measures for Ten Test Data Sets.

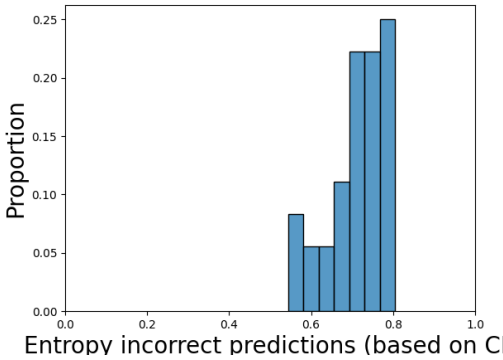
| **Accuracy Measure** | **Average** |
|---|---|
| *Mean Accuracy* | 78.84 % |
| *Mode Accuracy* | 78.59 % |
| *Median Accuracy* | 78.91 % |
| *AUC Accuracy* | 78.92 % |
| *95% CI Accuracy* | 92.57 % |
| *Percentage of the 95% CI correct predictions that contain 0.5* | 39.15 % |
| *Percentage of the 95% CI correct predictions that are either side of 0.5* | 60.85 % |
| *F-Score* | 0.623 |




(a) Histogram of the Entropy for All Test Predictions.



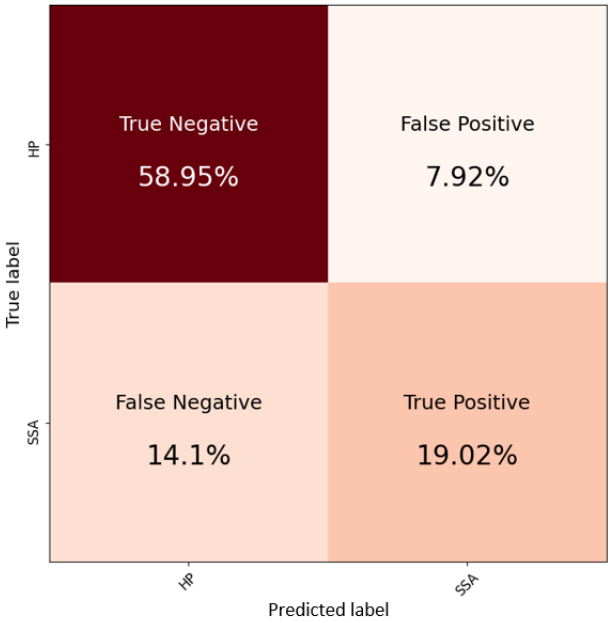

(b) Histogram of the Entropy for Rest Predictions where the Model made a Correct Prediction.



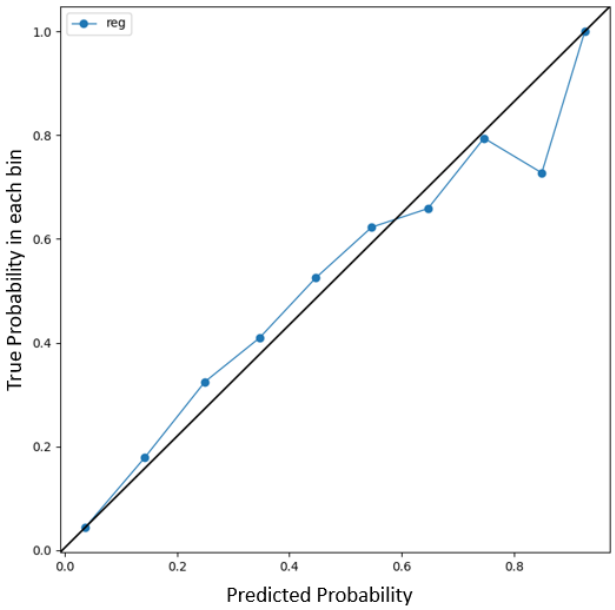

(c) Histogram of the Entropy for Test Predictions where the Model made an Incorrect Prediction.

Figure 5.3: Entropy Plots for Cancer Diagnosis Example

We can further assess model performance by assessing the behaviour of our model based on pathologist agreement levels on the entire test set. This is often referred to as *inter-rater* or *inter-annotator agreement*. When modelling data that has some inter-rater disagreement, it is important to have a well-calibrated model and a model that reflects any disagreement [43, 44]. In our case, we have a reasonably well-calibrated model, as shown in Figure 5.4b. If an analyst's model is poorly calibrated, they can use methods like label smoothing to help calibrate it [43, 45, 44]. Additionally, we need to ensure that our uncertainty estimates reflect the varying levels of inter-rater agree-




(a) Confusion Matrix




(b) Calibration Plot

Figure 5.4: Model Diagnostic Plots for Cancer Diagnosis Example

ment. To do this, we calculate the entropy of the probability samples for each observation in the set and also check whether its credible interval is centred. Observations are grouped into an "Agreement Level" category based on the number of pathologists who agreed on the diagnosis. If all pathologists agreed, the case is classified as "Full Agreement," if one pathologist gave a differing diagnosis, it is classified as "One Opposing," and so on. The results of this

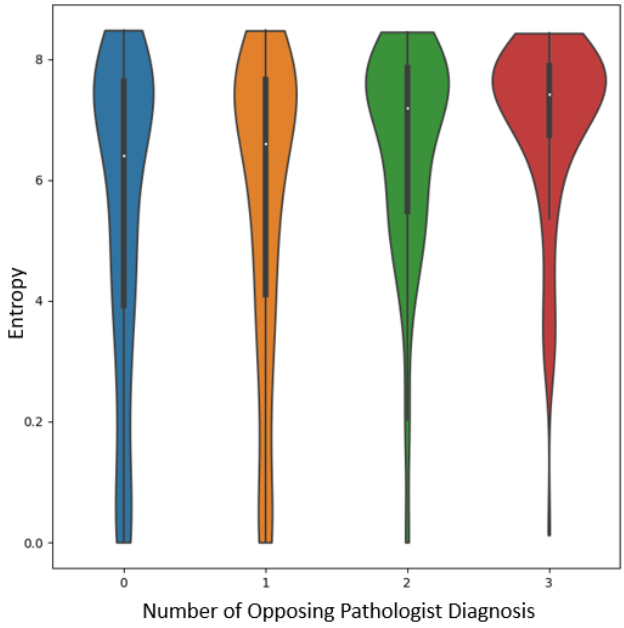
analysis are presented in Table 5.3. Our findings indicate that the mean entropy increases as the number of opposing pathologists increases, indicating that overall there is greater uncertainty in the estimated distributions when the pathologist group is uncertain. Furthermore, in most cases, the number of credible intervals that are centred increases with the level of disagreement among pathologists. This is a strong indication that the model is in line with the agreement levels of pathologists, exhibiting more uncertainty in cases where the group of pathologists exhibits greater disagreement. Figure 5.5a provides further evidence of this. Violin plots are useful when displaying this type of data as they easily display the summary statistics of the data (median: white dot, interquartile range: thick black bar) while also showing the shape of the density [46] (Figure 5.5b and 5.5a). We can see that for "Full Agreement" and "One Opposing", we have thicker tails (values closer to zero) in the plot of all test set entropy values, meaning that for these classes, the model is making more certain decisions; this is reduced for "Two Opposing" and "Three Opposing" where the tail is thinner and more values are in the higher end. These results show that overall the model is exhibiting conservative behaviour, which is appropriate when eliciting uncertainty for critical decision making, as discussed in Section 5.2.

The entropy we have considered so far has been the entropy of the whole distribution; we shall call this *distribution entropy*. Distribution entropy is important in assessing the width and height of the elicited distribution. However, there is another common entropy metric that is used for a categorical probability distribution, which can be computed by taking the mean of the sampled probabilities $p_i$ as an estimate of the probability that the (Bernoulli) event will occur, and the mean of $1 - p_i$ as the estimated probability that it will not occur. If an event has a mean probability of occurring, $q$, equal to 0.5, then we can assume that the model is uncertain whether the event will occur (high entropy value). However, if $q$ is 0.9, then we can assume that it is

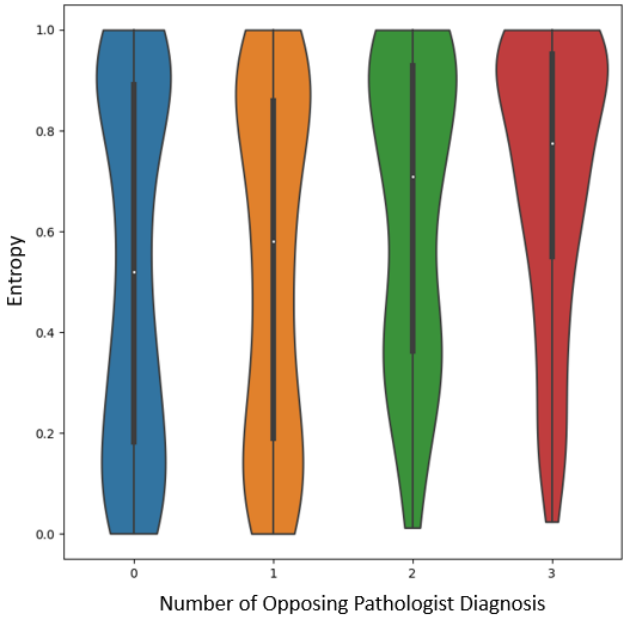
fairly certain the event will occur (low entropy value). We will call this *point estimate entropy*. This information is also captured in stating if the credible interval contains 0.5 (Table 5.3), but we can get a more specific view by looking at the point estimate entropy. We can assess the point estimate entropy for observations in each agreement level by taking the mean of the elicited distributions (Figure 5.5b). Figure 5.5b reiterates what we saw in Figure 5.5a, with the full agreement and the one opposing group having more values closer to zero than the other groups. Also, the interquartile range gets smaller, and the median of the violin plot increases as the number of opposing pathologists increases. Overall, our model is aligned with differing agreement levels.

Table 5.3: Diagnostics for Differing Pathologist Diagnoses

| **Agreement Level** | **Mean Entropy** | **Percentage of centred CI's** |
|---|---|---|
| *Full Agreement* | 0.6314 | 36.41 % |
| *One Opposing* | 0.6433 | 32.47% |
| *Two Opposing* | 0.7027 | 47.97% |
| *Three Opposing* | 0.7110 | 52.08% |




(a) Distribution entropy of test dataset points by agreement level.




(b) Mean point estimate entropy of test dataset points by agreement level.

Figure 5.5: Entropy Plots of the Elicited Distributions of all Test Points. Split by Agreement Levels.

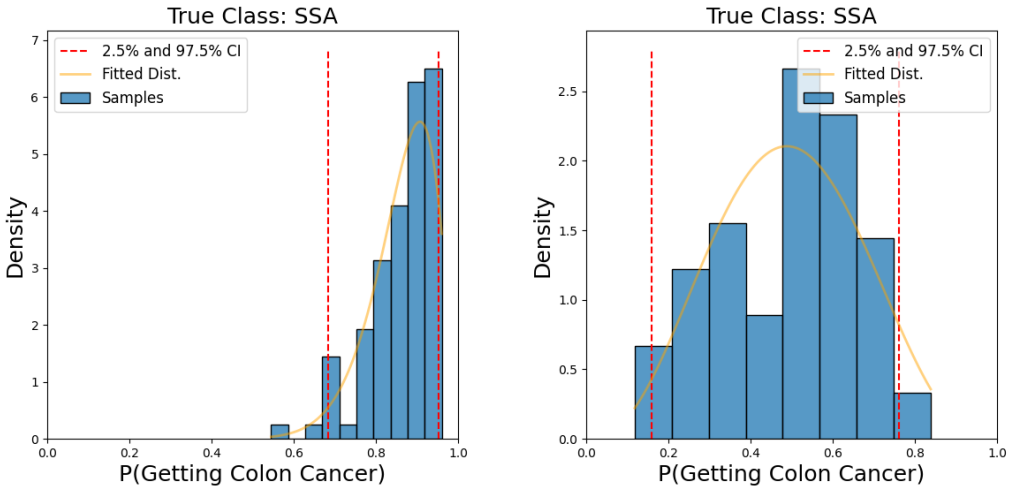
### 5.4.4 Elicited Distribution

After successfully training our model, we can now elicit distributions for specific individuals. This is done by inputting an image into the trained model 100 times, with dropout layers activated, to obtain a sample of 100 probabilities. The Method of Moments [47] is used to fit a beta distribution to the samples.

To show examples of elicited distributions, we chose four individuals from the data set and excluded them from any of the training and test data sets. Two of them were diagnosed with SSA, while the other two had HP. Images from each of these individuals were passed through the trained model, and distributions were obtained. We further evaluated our model's performance by selecting two individuals for each diagnosis based on the pathologists' agreement. One individual had a high number of pathologists agreeing on the diagnosis (seven out of seven), while the other had a low agreement (four out of seven). For individuals with high agreement, we expect the elicited distributions to resemble a certain expert decision, with a narrower distribution on either side of 0.5. Conversely, for individuals with low agreement, we expected our distributions to be wider and more centred around 0.5.

For Individual A, who was given the same diagnosis by all seven pathologists, our model elicited a $Beta(15.718, 2.502)$, and the probability distribution was left-skewed and relatively narrow (Figure 5.6a). In contrast, Individual B, given the same diagnosis by only four out of seven pathologists, had a more centred and wider distribution (fitted $Beta(3.656, 3.784)$). These results show that our model's predictions for SSA diagnosis align with pathologists' uncertainty in diagnoses. For patients given the HP diagnosis, there are similar results, albeit perhaps exhibiting slightly more uncertainty than for individuals with the SSA diagnosis. The distribution elicited for Individual C (fitted $Beta(3.953, 7.450)$) is less central than the elicited distribution for Individual D (fitted $Beta(4.591, 5.174)$), with the Individual C distribution being slightly

left skewed. Both distributions have similar widths (Figure 5.7).



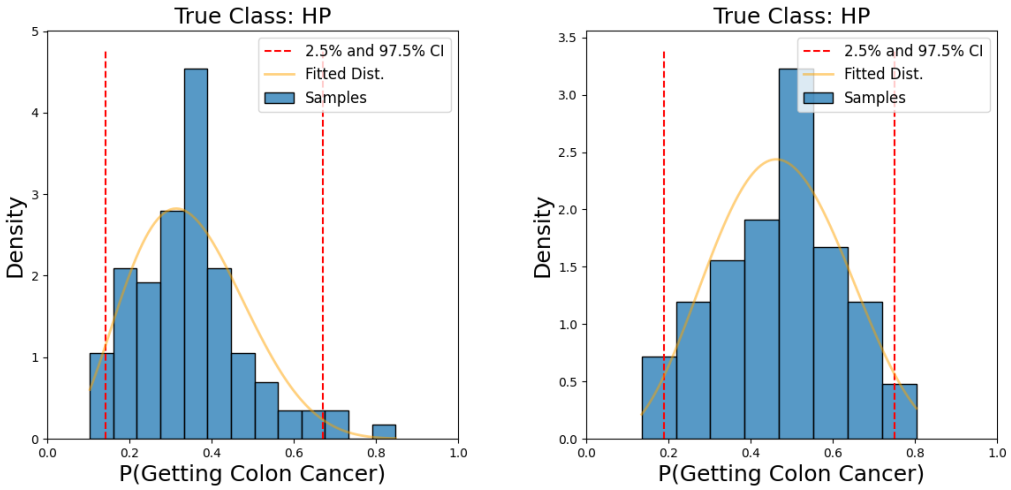

(a) Individual A: diagnosed with SSA where seven out of seven pathologists diagnosed SSA

(b) Individual B: diagnosed with SSA where four out of seven pathologists diagnosed SSA

Figure 5.6: Elicited Prior Probability Distributions for Individuals Diagnosed with SSA

## 5.5 Conclusions and Future Research

This article demonstrates the usage of the method for prior elicitation presented in [1] in the wider context of eliciting expert uncertainty, particularly for decision-making tasks with complex data like reports and images. We illustrate how by introducing probabilistic features in deep learning models, we can use records of expert decision-making to elicit expert uncertainty. In particular, applying dropout layers to a deep learning model is a simple yet effective method to obtain a probability distribution that encapsulates experts' uncertainty. As an example, we use dropout to elicit a distribution for estimating the risk of cancer in a patient's future. This example showcases how deep learning models can imitate decision making and capture experts' uncertainty. The dropout model performs well by eliciting both certain and uncertain distributions that align with the experts' uncertainty. The goal is for these models to conservatively capture expert uncertainty in the elicited




(a) Individual C: diagnosed with HP where seven out of seven pathologists diagnosed HP

(b) Individual D: diagnosed with HP where four out of seven pathologists diagnosed HP

Figure 5.7: Elicited Prior Probability Distributions for Individuals Diagnosed with HP

distributions.

Our example highlights the simplicity of using historical decision making data for uncertainty elicitation in healthcare. Medical literature typically uses expert elicitation methods that involve interviewing experts about distribution probabilities and quantiles [48, 49]. Analysts often set up the elicitation process using established protocols such as the SHELF protocol [50, 48, 51]. These methods tend to take a lot of time to produce sufficient results, as a considerable amount of preparatory work must be undertaken before the interviews with experts can proceed [51, 52]. Getting accurate results from interviews takes time and caution. First, an expert must be trained in probability theory, and then time must be set aside for experts to be interviewed and followed up with if required [51]. Then analysts must take precautions when asking questions, to ensure the elicitation process does not induce bias [49]. Modelling decision making data to elicit distributions requires little preparatory work, only for data to be collected and models to be built. It also does not induce bias through the elicitation process itself. This provides a simpler

option for implementation compared to typical methods. It is worth noting when considering applications to healthcare, the field of knowledge elicitation would benefit from more research into comparing currently available techniques. It would be useful to see the exact differences (if any) in distributions elicited through other methods and those elicited from expert decision making.

Expert knowledge elicitation is often a key component in critical and strategic decision-making. As such, our proposed deep learning based expert elicitation approach can be used in decision-making approaches based on decision theory, risk analysis or adversarial risk analysis (ARA) [53]. This is especially so where the decision-making is repetitive and frequent, yet each case is unique in its own way and the decision has important implications. ARA, in particular, is based on game Theory principles and aims to find the optimal actions for a defender, while taking into account prior uncertainty surrounding an offender; any uncertainty must be elicited from experts before ARA can be applied. The method presented in this paper can be used to elicit these distribution for ARA models for repetitive decision-making as highlighted in [54]. Yet, the potential applications are beyond those in the justice and medical diagnosis fields and also includes applications related to actuarial and security risks, for example. Future research should be undertaken to show how the proposed expert elicitation approach can be used in such models.

Lastly, it is important to note that deep learning is a proliferating field with many different research avenues. Analysts should expect research to significantly improve models in the near future, which will make applying deep learning models to knowledge elicitation tasks easier.

# References

[1] Julia R Falconer, Eibe Frank, Devon LL Polaschek, and Chaitanya Joshi. Eliciting informative priors by modelling expert decision making. *arXiv preprint arXiv:2307.07098*, 2023.

[2] M Granger Morgan. Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National academy of Sciences*, 111(20):7176–7184, 2014.

[3] Anthony O'Hagan, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons, 2006.

[4] Marcelo Hartmann, Georgi Agiashvili, Paul Bürkner, and Arto Klami. Flexible prior elicitation via the prior predictive distribution. In *Conference on Uncertainty in Artificial Intelligence*, pages 1129–1138. PMLR, 2020.

[5] Owen Thomas, Henri Pesonen, and Jukka Corander. Probabilistic elicitation of expert knowledge through assessment of computer simulations. *arXiv preprint arXiv:2002.10902*, 2020.

[6] Christopher J Casement and David J Kahle. Graphical prior elicitation in univariate models. *Communications in Statistics-Simulation and Computation*, 47(10):2906–2924, 2018.

[7] Robert L Winkler. The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical association*, 62(319):776–800, 1967.

[8] Julia R Falconer, Eibe Frank, Devon LL Polaschek, and Chaitanya Joshi. Methods for eliciting informative prior distributions: A critical review. *Decision Analysis*, 19(3):189–204, 2022.

[9] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974.

[10] Robert T Eckenrode. Weighting multiple criteria. *Management science*, 12(3):180–192, 1965.

[11] Ward Edwards and F Hutton Barron. Smarts and smarter: Improved simple methods for multiattribute utility measurement. *Organizational behavior and human decision processes*, 60(3):306–325, 1994.

[12] Chen Wang and Vicki M Bier. Expert elicitation of adversary preferences using ordinal judgments. *Operations Research*, 61(2):372–385, 2013.

[13] Robert L Winkler. The quantification of judgment: Some methodological suggestions. *Journal of the American Statistical Association*, 62(320):1105–1120, 1967.

[14] Mohamad Amin Pourhoseingholi, Ahmad Reza Baghestani, and Mohsen Vahedi. How to control confounding effects by statistical analysis. *Gastroenterology and hepatology from bed to bench*, 5(2):79, 2012.

[15] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

[16] Znaonui Liang, Gang Zhang, Jimmy Xiangji Huang, and Qmming Vivian Hu. Deep learning for healthcare decision making with emrs. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 556–559. IEEE, 2014.

[17] Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Vaickus, Charles Brown, Michael Baker, Naofumi Tomita, Lorenzo Torresani, et al. A petri dish for histopathology image analysis. In *Artificial Intelligence in Medicine: 19th International Conference on Artificial*

*Intelligence in Medicine, AIME 2021, Virtual Event, June 15–18, 2021, Proceedings*, pages 11–24. Springer, 2021.

[18] Fabio Campos, Andre Neves, and Fernando M Campello de Souza. Decision making under subjective uncertainty. In *2007 IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making*, pages 85–90. IEEE, 2007.

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

[20] Stephen C Hora. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54(2-3):217–223, 1996.

[21] Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D Phillips. Calibration of probabilities: The state of the art. *Decision making and change in human affairs*, pages 275–324, 1977.

[22] Tommi Perälä, Jarno Vanhatalo, Anna Chrysafi, et al. Calibrating expert assessments using hierarchical gaussian process models. *Bayesian analysis*, 15(4):1251–1280, 2020.

[23] J Edward Russo and Paul JH Schoemaker. Managing overconfidence. *Sloan management review*, 33(2):7–17, 1992.

[24] Steffen Andersen, John Fountain, Glenn W Harrison, and E Elisabet Rutström. Estimating subjective probabilities. *Journal of Risk and Uncertainty*, 48:207–229, 2014.

[25] Keith J Beven, Willy P Aspinall, Paul D Bates, Edoardo Borgomeo, Katsuichiro Goda, Jim W Hall, Trevor Page, Jeremy C Phillips, Michael Simpson, Paul J Smith, et al. Epistemic uncertainties and natural hazard risk assessment–part 2: What should constitute good practice? *Natural Hazards and Earth System Sciences*, 18(10):2769–2783, 2018.

[26] Jean-Charles Pomerol. Artificial intelligence and human decision making. *European Journal of Operational Research*, 99(1):3–25, 1997.

[27] Kevin Gurney. *An introduction to neural networks*. CRC press, 2018.

[28] Yann LeCun et al. Generalization and network design strategies. *Connectionism in perspective*, 19:143–155, 1989.

[29] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[30] 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.

[31] Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. Hands-on Bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022.

[32] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[33] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[34] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015.

[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] Rohit Makkar, Rish K Pai, and Carol A Burke. Sessile serrated polyps: cancer risk and appropriate surveillance. *Cleve Clin J Med*, 79(12):865–71, 2012.

[37] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 186–195. Springer, 2021.

[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[39] Christopher M Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.

[40] Yutaka Sasaki et al. The truth of the f-measure. *Teach tutor mater*, 1(5):1–5, 2007.

[41] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[42] David JC MacKay, David JC Mac Kay, et al. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[43] Martin Holm Jensen, Dan Richter Jørgensen, Raluca Jalaboi, Mads Eiler Hansen, and Martin Aastrup Olsen. Improving uncertainty estimation in convolutional neural networks using inter-rater agreement. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 540–548. Springer, 2019.

[44] Andreanne Lemay, Charley Gros, Enamundram Naga Karthik, and Julien Cohen-Adad. Label fusion and training methods for reliable representation of inter-rater uncertainty. *arXiv preprint arXiv:2202.07550*, 2022.

[45] Mobarakol Islam and Ben Glocker. Spatially varying label smoothing: Capturing uncertainty from expert annotations. In *Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27*, pages 677–688. Springer, 2021.

[46] Jerry L Hintze and Ray D Nelson. Violin plots: a box plot-density trace synergism. *The American Statistician*, 52(2):181–184, 1998.

[47] Karl Pearson. Method of moments and method of maximum likelihood. *Biometrika*, 28(1/2):34–59, 1936.

[48] Danila Azzolina, Paola Berchialla, Dario Gregori, and Ileana Baldi. Prior elicitation for use in clinical trial design and analysis: a literature review. *International journal of environmental research and public health*, 18(4):1833, 2021.

[49] Sindhu R Johnson, George A Tomlinson, Gillian A Hawker, John T Granton, Haddas A Grosbein, and Brian M Feldman. A valid and reliable belief elicitation method for Bayesian priors. *Journal of clinical epidemiology*, 63(4):370–383, 2010.

[50] John Paul Gosling. Shelf: the sheffield elicitation framework. *Elicitation: The science and art of structuring judgement*, pages 61–93, 2018.

[51] Sara Graziadio and Kevin J Wilson. Uncertainty representation for early phase clinical test evaluations: a case study. *arXiv preprint arXiv:2005.10011*, 2020.

[52] Sabrina H Rossi, Christopher Blick, Paul Nathan, David Nicol, Grant D Stewart, and Edward CF Wilson. Expert elicitation to inform a cost-

effectiveness analysis of screening for renal cancer. *Value in Health*, 22(9):981–987, 2019.

[53] David Rios Insua, Jesus Rios, and David Banks. Adversarial risk analysis. *Journal of the American Statistical Association*, 104(486):841–854, 2009.

[54] Chaitanya Joshi, Charnè Nel, Javier Cano, and Devon L.L. Polascheck. Parole board decision-making using adversarial risk analysis. *(under review)*, 2023.

# Chapter 6

# Conclusion

Expert knowledge elicitation can be used for many tasks, including eliciting distributions that capture uncertainty and prior distributions for Bayesian inference. This research aims to rejuvenate the field of uncertainty elicitation by highlighting the limitations of current techniques and providing modern solutions.

## 6.1 Discussion

The field of prior elicitation previously directed a lot of attention to one method, Direct Interrogation. Although this method has proven itself credible in many applications, there are times when it is not appropriate. The key shortcoming is that it requires experts to have sufficient probability knowledge to elicit their prior, and the analysts must directly interact with experts to produce this prior. Other methods, such as *Indirect Interrogation*, eradicate the requirement of experts needing knowledge of probability theory, however, they do not place importance on experts to be accurate in their responses. Dealing with experts to elicit uncertainty also requires analysts to address cognitive biases. Using methods that limit the biases or help an analyst to address them would be beneficial. Ideally, an analyst would want to reduce the dependency on experts, to obtain a distribution that captures uncertainty. However, meth-

ods that involve no experts and instead focus on eliciting distributions from data are impractical when no data is available.

This thesis introduced a method for prior elicitation that combines the benefits of using experts and using historical data while reducing the shortcomings of current methods. The basic structure of the method involves using probabilistic modelling to elicit uncertainty from an expert decision-making task. If there is a tabular dataset of appropriate structure that encompasses useful decision-making data and the decision outcome, an analyst can use Bayesian logistic regression to elicit a distribution that captures uncertainty from the dataset. Chapter 4 showed how this method can be used in practice with an example involving the parole board decision-making process. This data was used to elicit the uncertainty surrounding a prisoner recommitting a crime upon release from prison. Caution needs to be taken when selecting the best model for elicitation, and a conservative approach should be considered (explained in Chapter 5). Model accuracy and the appropriateness of the elicited distributions need to be weighed up. An analyst must also consider if there are any biases present in the decision-making process. Using tabular data, there is a simple way for us to explore if there are any biases present by removing the corresponding variables from model building. This allows an analyst to easily see the effects of the variable on the elicited distributions by comparing two distributions, one elicited with the corresponding variable in the model and one elicited without the corresponding variable. For the parole board example, it was shown that the ethnicity variable was having an impact. Determining whether or not ethnicity is actually inducing bias is a little more challenging with tabular data, as ethnicity could be a confounding variable, meaning that the variable could be capturing information that is not explicitly represented in the dataset. For ethnicity, this could be socio-economic circumstances.

Tabular data is limited in the information it can provide to the model to

elicit distributions. Variables that may be considered biased could, in fact, be confounding variables. However, modelling all the information used to make the decision will not exhibit this problem. The issue is that the information used by experts to make decisions is usually not in tabular form and instead may consist of images and reports. Chapter 5 explored eliciting distributions that capture expert uncertainty from images using a deep learning approach. The key to using deep learning to elicit these distributions is to add some probabilistic features to the models. We advise using drop-out layers as they are simple to implement yet still produce good results. If the data is analysed by multiple experts and there is disagreement among them in making a decision, this can help determine how accurately the elicited distributions align with expert opinions, as shown in 5.

The field of expert uncertainty elicitation needs new methods to be explored, using modern statistical techniques. Past methods have many limitations and are not appropriate for many tasks. The new method outlined in this thesis, eliciting uncertainty from expert decision-making, combats many of the challenges faced by existing methods and provides new ways to look at bias in decision-making. Although the proposed method has some limitations, particularly when decision-making data is not available, it provides an attractive option in many practical applications. Alternative solutions, such as direct and indirect interrogation, are outlined in Chapter 3. It is acknowledged that deep learning models have their own limitations, such as processing power, but given the rapid advancements in this field, we are optimistic that these constraints will be addressed in the near future.

Modern literature on prior elicitation often follows Garthwaite et al's [1] definition of elicitation, "the process of formulating a person's knowledge and beliefs about one or more uncertain quantities into a (joint) probability distribution for those quantities". This thesis aims to bring prior elicitation to a

wider audience and explain in detail all the methods for obtaining informative priors that an analyst can use, for this reason, the term elicitation is taken from its root meaning, to obtain something [2]. Hence, the definition of prior elicitation is "the process of obtaining knowledge from a source to form a prior distribution". As discussed throughout, the method in this thesis aims to form a prior distribution from expert decision-making through probabilistic modelling. Although standard probabilistic models are used, this method is an expert knowledge elicitation/ expert prior elicitation technique because expert decision-making is being modelled to form a distribution. This means experts are inputting their own subjective opinion to form their decisions, this information is modelled to infer a distribution that captures expert uncertainty.

## 6.2 Future Research

In this thesis, we propose a new method for eliciting expert uncertainty that can be extended to various tasks. Further research could apply this method to more complex situations than those discussed in this thesis, including those with a decision-making task that has non-binary or multinomial responses (briefly discussed in Chapter 4 Section 4.5) or those with multi-modal data. Multi-modal data, data that contains multiple formats such as images and reports, is commonly used for decision-making, and there already exist solutions in deep learning to model it [3]. It would be valuable to see how our proposed method can be extended to many different real-life situations.

We make a point of highlighting the limited amount of work done on comparing currently available elicitation methods. Current work [4, 5, 6] has focused on comparing methods of one type of elicitation technique, for example, interrogation methods [4]. In contrast, in this work, we have highlighted many different elicitation methods and also introduced a new method. Although theoretically, we know the limitations of all these methods, it would be beneficial to see what (if any) differences are in the elicited distributions from

each method in practice. Future research should focus on comparing a range of different elicitation methods. This will help practitioners to know which method works best for their task and may even highlight further shortcomings of some methods.

The parole board example in Chapter 4 presents a unique situation for prior elicitation tasks, where data can be observed in the future and may lead to a system change. For the parole board example, if we have elicited a prior distribution for a specific prisoner recommitting a crime upon release from the parole board and said prisoner is released, we can observe whether or not they do commit a crime. Future research could be undertaken to see if the model used to elicit the distribution can be calibrated to account for the now observed data. Calibration techniques that use past estimates given by the expert where the true value is now known have been explored [7, 8]. It would be beneficial to investigate if a calibration technique can be applied to the method introduced in this thesis. It is hoped that incorporating this feature into the model will allow future elicited distributions for other prisoners to be calibrated to the observed events and reduce the effects of expert bias.

# References

[1] Paul H Garthwaite, Joseph B Kadane, and Anthony O'Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–701, 2005.

[2] Cambridge University Press. *Cambridge Academic Content Dictionary*. Cambridge Academic Content Dictionary. Cambridge University Press, 2008.

[3] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011.

[4] Joseph Kadane and Lara J Wolfson. Experiences in elicitation: [read before the royal statistical society at a meeting on'elicitation 'on wednesday, april 16th, 1997, the president, professor afm smith in the chair]. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):3–19, 1998.

[5] Ali E Abbas, David V Budescu, Hsiu-Ting Yu, and Ryan Haggerty. A comparison of two probability encoding methods: Fixed probability vs. fixed variable values. *Decision Analysis*, 5(4):190–202, 2008.

[6] Cameron J Williams, Kevin J Wilson, and Nina Wilson. A comparison of prior elicitation aggregation using the classical method and shelf. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2021.

[7] Tommi Perälä, Jarno Vanhatalo, Anna Chrysafi, et al. Calibrating expert assessments using hierarchical gaussian process models. *Bayesian analysis*, 15(4):1251–1280, 2020.

[8] Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D Phillips. Calibration of probabilities: The state of the art. *Decision making and change in human affairs*, pages 275–324, 1977.

# Appendix A

# Co-Authorship Forms

The co-authorship forms related to the three articles included in this thesis are provided on the following pages

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

**Postgraduate Studies Office**
Student and Academic Services Division
*Wahanga Ratonga Matauranga Akonga*
The University of Waikato
Private Bag 3105
Hamilton 3240, New Zealand
Phone +64 7 838 4439
Website: http://www.waikato.ac.nz/sasd/postgraduate/

# Co-Authorship Form

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. **Please include one copy of this form for each co-authored work**. Completed forms should be included in your appendices for all the copies of your thesis submitted for examination and library deposit (including digital deposit).

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter Three: Methods for eliciting informative prior distributions: A critical review.
Published to Decision Analysis

| | |
|---|---|
| Nature of contribution by PhD candidate | Conceptualisation, Writing and Reviewing |
| Extent of contribution by PhD candidate (%) | 70 |

## CO-AUTHORS

| Name | Nature of Contribution |
|---|---|
| Eibe Frank | Provided advice throughout the project on computational aspects, deep learning and writing |
| Devon Polaschek | Provided advice throughout the project on bias, psychology and writing |
| Chaitanya Joshi | Provided advice throughout the project on Conceptualisation, Bayesian formulation, content and writing. |
| | |
| | |
| | |

## Certification by Co-Authors

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and

| Name | Signature | Date |
|---|---|---|
| Eibe Frank | Eibe Frank | 22 August 2023 |
| Devon Polaschek | Devon Polaschek | August 22, 2023 |
| Chaitanya Joshi | Joshi | 22-Aug-2023 |
| | | |
| | | |
| | | |

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

**Postgraduate Studies Office**
Student and Academic Services Division
*Wahanga Ratonga Matauranga Akonga*
The University of Waikato
Private Bag 3105
Hamilton 3240, New Zealand
Phone +64 7 838 4439
Website: http://www.waikato.ac.nz/sasd/postgraduate/

# Co-Authorship Form

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. **Please include one copy of this form for each co-authored work**. Completed forms should be included in your appendices for all the copies of your thesis submitted for examination and library deposit (including digital deposit).

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter Four: Eliciting Informative Priors by Modelling Expert Decision Making.
Submitted to Decision Analysis.

| | |
|---|---|
| Nature of contribution by PhD candidate | Conceptualisation, Modelling, Computation, Writing and Reviewing |
| Extent of contribution by PhD candidate (%) | 70 |

## CO-AUTHORS

| Name | Nature of Contribution |
|---|---|
| Elibe Frank | Provided advice throughout the project on computational aspects, deep learning and writing |
| Devon Polaschek | Provided advice throughout the project on bias, psychology and writing |
| Chaitanya Joshi | Provided advice throughout the project on Conceptualisation, Bayesian formulation, content and writing |
| | |
| | |
| | |

## Certification by Co-Authors

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and

| Name | Signature | Date |
|---|---|---|
| Eibe Frank | Eibe Frank | 22 August 2023 |
| Devon Polaschek | Devon Polaschek | August 22, 2023 |
| Chaitanya Joshi | Joshi | 22-Aug-2023 |
| | | |
| | | |
| | | |

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

# Co-Authorship Form

**Postgraduate Studies Office**
Student and Academic Services Division
*Wahanga Ratonga Matauranga Akonga*
The University of Waikato
Private Bag 3105
Hamilton 3240, New Zealand
Phone +64 7 838 4439
Website: http://www.waikato.ac.nz/sasd/postgraduate/

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. **Please include one copy of this form for each co-authored work**. Completed forms should be included in your appendices for all the copies of your thesis submitted for examination and library deposit (including digital deposit).

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter Five: Utilising Deep Learning to Elicit Expert Uncertainty.
Submitted to American Statistician

| | |
|---|---|
| Nature of contribution by PhD candidate | Conceptualisation, Modelling, Computation, Writing and Reviewing |
| Extent of contribution by PhD candidate (%) | 70 |

## CO-AUTHORS

| Name | Nature of Contribution |
|---|---|
| Elibe Frank | Provided advice throughout the project on computational aspects, deep learning and writing |
| Devon Polaschek | Provided advice throughout the project on bias, psychology and writing |
| Chaitanya Joshi | Provided advice throughout the project on Conceptualisation, Bayesian formulation, content and writing |
| | |
| | |
| | |

## Certification by Co-Authors

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and

| Name | Signature | Date |
|---|---|---|
| Eibe Frank | Eibe Frank | 22 August 2023 |
| Devon Polaschek | Devon Polaschek | August 22, 2023 |
| Chaitanya Joshi | Joshi | 22-Aug-2023 |
| | | |
| | | |
| | | |