Submitted to *Decision Analysis*
manuscript

Eliciting Informative Priors by Modelling Expert Decision Making

Julia R. Falconer
Department of Mathematics, University of Waikato, Hamilton New Zealand,
New Zealand Institute of Security and Crime Science, University of Waikato, Hamilton, New Zealand
jrg22@students.waikato.ac.nz

Eibe Frank
Department of Computer Science, University of Waikato, Hamilton New Zealand,
eibe@waikato.ac.nz,

Devon L. L. Polaschek
School of Psychology, University of Waikato, Hamilton New Zealand,
New Zealand Institute of Security and Crime Science, University of Waikato, Hamilton, New Zealand
polascde@waikato.ac.nz,

Chaitanya Joshi
Department of Statistics, University of Auckland, Auckland, New Zealand
chaitanya.joshi@auckland.ac.nz,

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.

Key words: Bayesian methods, prior elicitation, subjective, prior distribution, uncertainty

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, θ:

$$p(\theta|y) \propto p(\theta)p(y|\theta), \tag{1}$$

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 (Zyphur and Oswald 2015).

Table 1 Definitions expanded from a table from Falconer et al. (2022)

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 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 Falconer et al. (2022), which assigns methods to three categories; 1) Direct Interrogation Methods, 2) Indirect Interrogation Methods, and 3) Graphical/ Visual Methods. Direct Interrogation methods (O'Hagan et al. 2006, Galway 2007, Jenkinson 2005) 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 (O'Hagan et al. 2006, Thomas et al. 2020, Casement and Kahle 2018), but this can prove difficult and create inaccurate prior distributions (Kadane and Wolfson 1998, Wang and Bier 2013). This issue can also be seen in some graphical/visual methods (Falconer et al. 2022). 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 (Winkler 1967) and getting the expert to rank the likelihood of events (Eckenrode 1965, Edwards and Barron 1994, Wang and Bier 2013). As highlighted in Falconer et al. (2022), 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 (Kahneman et al. 1982), 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 (Kahneman et al. 1982), 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 (O'Hagan 2019). 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* (Janis 1983). 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 (Janis 1983). Instead of having experts reach a consensus, some methods allow analysts to combine experts' individual priors mathematically (O'Hagan et al. 2006) 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 (Press 2009)), 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 (Belenguer 2022). A famous example is the Correctional Offender Management Profiling for Alternative Sanction, COMPAS (Angwin et al. 2016). 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 (Angwin et al. 2016). This was because the model had learned from historic discriminatory court cases and enhanced the prejudices in the judicial system (Belenguer 2022). 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 (Dastin 2018, Belenguer 2022). This was because the model learned patterns from historic data where women were not hired for positions at tech companies (Belenguer 2022). 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 (Jargowsky 2005). 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 (O'Hagan 2019, O'Hagan et al. 2006).

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 affecting 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 2. We discuss ways to assess model behaviour in Section 3, with Section 4 outlining a simple example application. Finally, we close in Section 5 with a summary

of conclusions and further work.

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.

$$P(Y^*|X^*,\mathbf{Y},\mathbf{X}) = \int P(Y^*|X^*,\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{Y},\mathbf{X})\, d\boldsymbol{\theta}. \tag{2}$$

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,

$$A_i|X_i \overset{\text{d}}{\approx} Y_i|X_i. \tag{3}$$

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 (2). That is,

$$\pi(A^*|X^*) \approx P(Y^*|X^*,\mathbf{Y},\mathbf{X}). \tag{4}$$

See the accompanying influence diagram (Figure 1) that depicts the relationship between the variables.

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

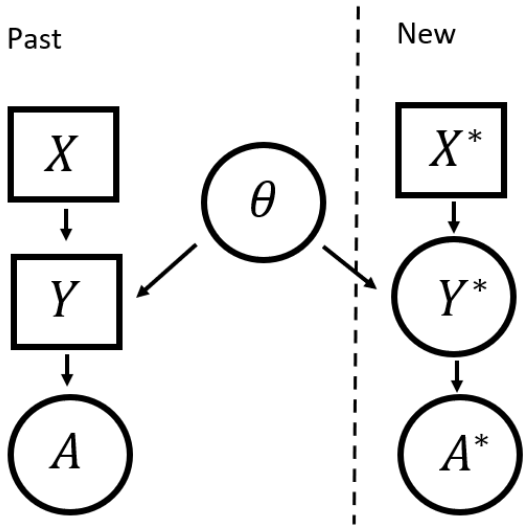


Figure 1 Influence Diagram for Eliciting Prior Distributions from Expert Decision Making

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 accurately 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 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 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,

$$g(p_i) = \theta_0 + \theta_1 x_1 i + ...$$

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

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

$$\Rightarrow p_i = \frac{exp(\theta_0 + \theta_1 x_i)}{1 + exp(\theta_0 + \theta_1 x_i)} \tag{5}$$

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 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 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 (Tollenaar and van der Heijden 2013, Caulkins et al. 1996, de la Cruz et al. 2021, Schmidt and Witte 1989)). 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.

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 labels for model

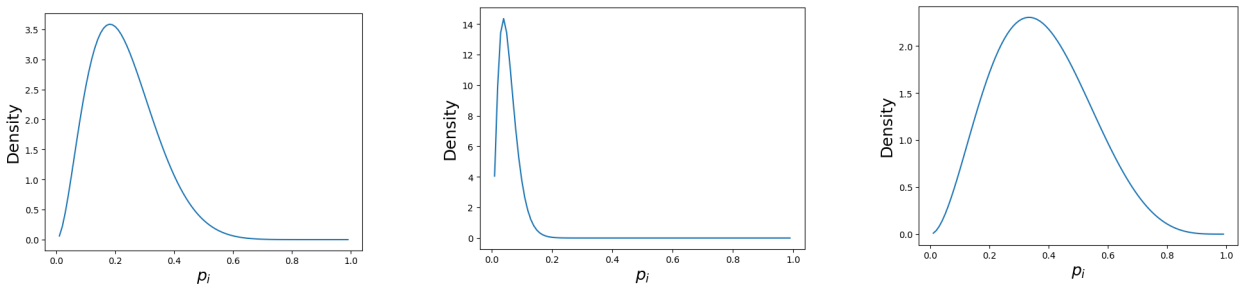


Figure 2 **Distributions of p_i for three different individuals that would obtain the same label assigned based on mean probability prediction.**

accuracy may not give a fair representation of the variability of the distributions. For example, Figure 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.

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 2). 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. 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

Table 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 (Sasaki et al. 2007)	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. $$F = 2.\frac{specificity.sensitivity}{specificity + sensitivity}$$.
Confusion Matrix (Fawcett 2006)	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 (MacKay et al. 2003)	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.

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.

4.1. 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 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 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).

4.2. Model

We wish to model the Parole Board Decision (response variable) using all other variables as explanatory variables (Table 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

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

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

Table 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 (of Statistics 2011).

(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 6.

$$\begin{aligned} Decision_i = {} & \beta_0 + \beta_1 \times gender\_male_i + \beta_2 \times age_i + \beta_3 \times num\_years\_release_i + \beta_4 \times num\_years\_parole_i \\ & + \beta_5 \times crime\_count_i + \beta_6 \times agg\_min\_sent_i + \beta_7 \times agg\_max\_sent_i + \beta_8 \times eth\_hispanic_i \\ & + \beta_9 \times eth\_white_i + \beta_{10} \times eth\_other_i + \beta_{11} \times crime\_class\_B_i + \beta_{12} \times crime\_class\_C_i \end{aligned}$$

$$+ \beta_{13} \times crime\_class\_D_i + \beta_{14} \times crime\_class\_E_i + \beta_{15} \times crime\_conviction\_assault_i$$
$$+ \beta_{16} \times crime\_conviction\_burglary_i + \ldots \quad (6)$$
$$p_i = \frac{1}{1 + e^{Decision_i}}$$

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

4.3. Model Diagnostics

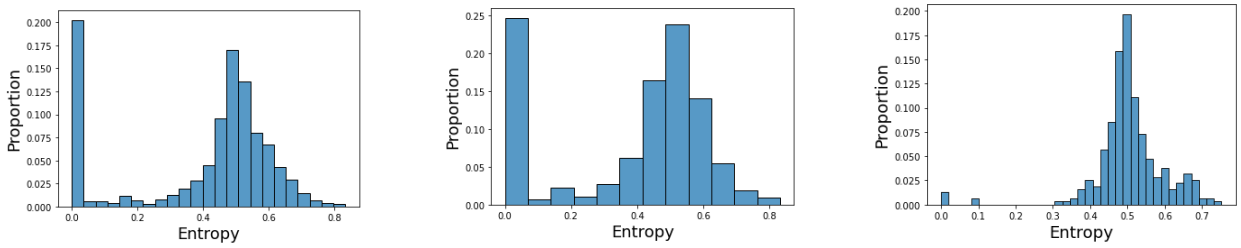
Accuracy readings were taken for the five different test sets and can be found in Table 4. The model obtains about 79% classification accuracy overall. The CI accuracy is approximately 84%, with 87% of the CIs being on either side of 0.5, showing that the model is making more certain predictions than 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 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 3b displays the entropy of correct predictions the model made and Figure 3c displays the entropy of incorrect predictions. We can see that for incorrect predictions the large peak at zero is not present (Figure 3c), whereas it is still present for correct predictions (Figure 3b) meaning our model is less certain with its predictions when it is incorrect. The model looks relatively well-calibrated to the data (Figure 4a). The confusion matrix (Figure 4b) shows that the model has a high true positive rate,

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

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 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



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

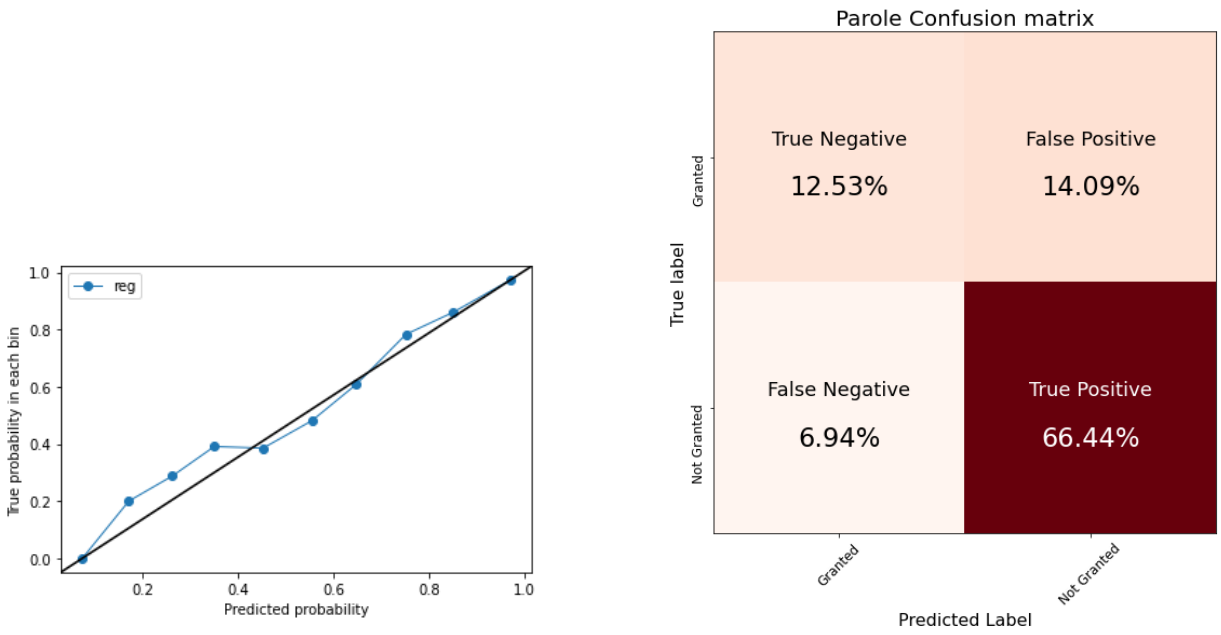
(b) Histogram of the entropy for test predictions where the model made a correct prediction.

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

Figure 3 Entropy Plots

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



(a) Calibration Plot (b) Confusion Matrix

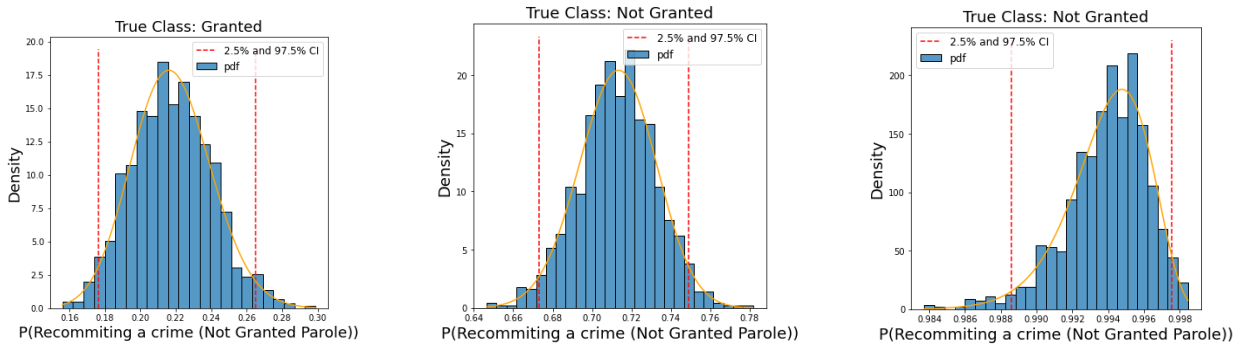
Figure 4 Other performance plots

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 (Massey Jr 1951).

Consider three prisoners: Prisoner 1, Prisoner 2 and Prisoner 3 (the prisoners' attributes are found in Table 5). The elicited prior distributions are shown in Figure 5. Prisoner 1 yielded a $Beta \sim (74.111, 266.202)$ prior distribution (Figure 5a). Prisoner 2 yielded a $Beta \sim (382.491, 154.224)$ prior distribution (Figure 5b). Prisoner 3 yielded a $Beta \sim (1181.395, 7.210)$ prior distribution (Figure 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 5 : The prisoners' attributes used in Example 1

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



(a) Prisoner 1 (b) Prisoner 2 (c) Prisoner 3

Figure 5 Prior distributions for three different prisoners

4.5. Influential Variables

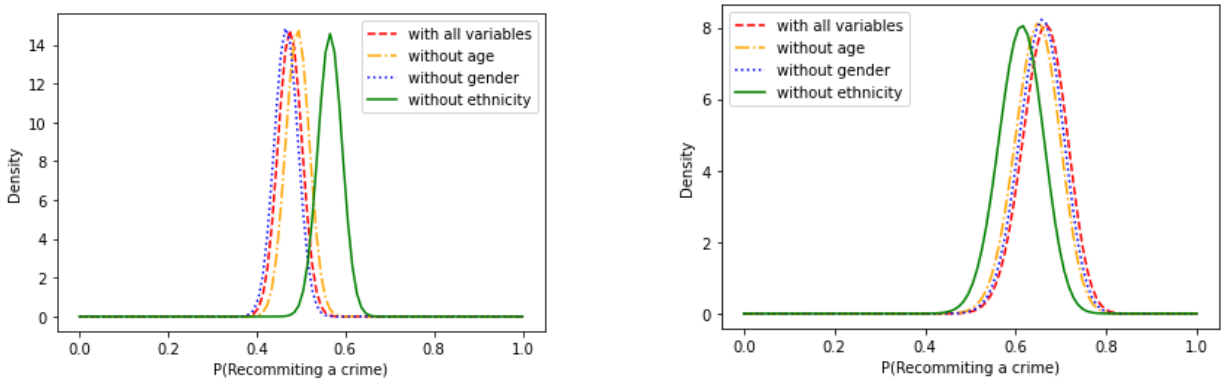
For this example, we have shown how an analyst can elicit a prior distribution from an expert decision-making process using tabular data. However, can an analyst trust that this elicited distribution is reliable? Can they trust the expert's decisions? Could some variables be wrongly influencing decisions? 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 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 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 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 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 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 (Association et al. 2017). 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 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

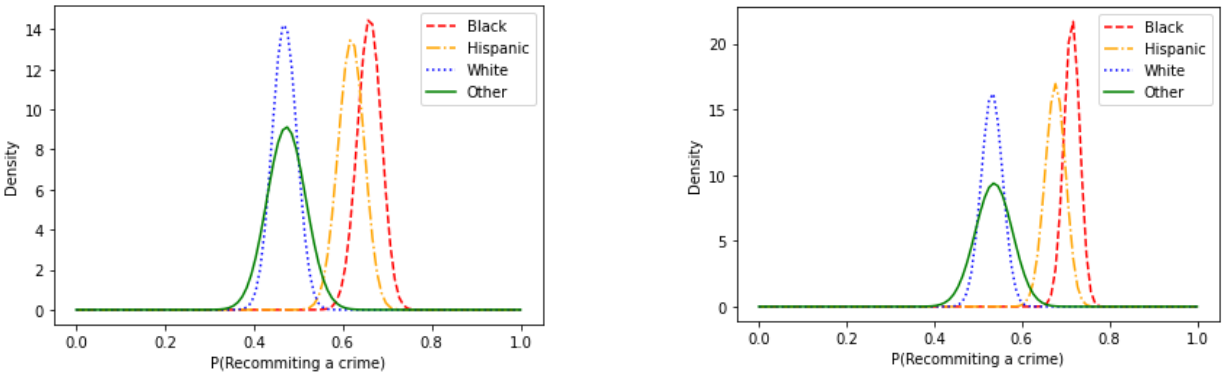


(a) Prisoner 4 (b) Prisoner 5

Figure 6 Elicited distributions for the four different models for different prisoners.

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 different results to those



(a) Prisoner 4 (b) Prisoner 5

Figure 7 **Elicited distributions for the same test point but with ethnicity variable changed**

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 (Perälä et al. 2020) and should be a continual path for future research (discussed further in Section 5).

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 θ, 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 2, 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 θ, 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 (Fisher and McEvoy (2022), O'brien and Dunson (2004)) to elicit distributions. This can be done in a similar way to that discussed in Section 2 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 Perälä et al. (2020)). By using a number of accuracy measures (discussed in Section 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

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias. *Ethics of data and analytics*, 254–264 (Auerbach Publications).

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

Belenguer L (2022) 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.

Casement CJ, Kahle DJ (2018) Graphical prior elicitation in univariate models. *Communications in Statistics-Simulation and Computation* 47(10):2906–2924.

Caulkins J, Cohen J, Gorr W, Wei J (1996) Predicting criminal recidivism: A comparison of neural network models with statistical methods. *Journal of Criminal Justice* 24(3):227–240.

Dastin J (2018) Amazon scraps secret ai recruiting tool that showed bias against women. *Ethics of data and analytics*, 296–299 (Auerbach Publications).

de la Cruz R, Padilla O, Valle MA, Ruz GA (2021) Modeling recidivism through bayesian regression models and deep neural networks. *Mathematics* 9(6):639.

Eckenrode RT (1965) Weighting multiple criteria. *Management science* 12(3):180–192.

Edwards W, Barron FH (1994) Smarts and smarter: Improved simple methods for multiattribute utility measurement. *Organizational behavior and human decision processes* 60(3):306–325.

Falconer JR, Frank E, Polaschek DL, Joshi C (2022) Methods for eliciting informative prior distributions: A critical review. *Decision Analysis* 19(3):189–204.

Fawcett T (2006) An introduction to roc analysis. *Pattern recognition letters* 27(8):861–874.

Fisher JD, McEvoy KR (2022) Bayesian multinomial logistic regression for numerous categories. *arXiv preprint arXiv:2208.14537* .

Galway LA (2007) Subjective probability distribution elicitation in cost risk analysis: A review .

Janis IL (1983) *Groupthink* (Houghton Mifflin Boston).

Jargowsky PA (2005) Omitted variable bias. *Encyclopedia of social measurement* 2:919–924.

Jenkinson D (2005) The elicitation of probabilities: A review of the statistical literature .

Kadane J, Wolfson LJ (1998) 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.

Kahneman D, Slovic SP, Slovic P, Tversky A (1982) *Judgment under uncertainty: Heuristics and biases* (Cambridge university press).

MacKay DJ, Mac Kay DJ, et al. (2003) *Information theory, inference and learning algorithms* (Cambridge university press).

Massey Jr FJ (1951) The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association* 46(253):68–78.

O'brien SM, Dunson DB (2004) Bayesian multivariate logistic regression. *Biometrics* 60(3):739–746.

of Statistics AB (2011) Australian and New Zealand Standard Offence Classification (ANZSOC). `https://www.abs.gov.au/statistics/classifications/`

`australian-and-new-zealand-standard-offence-classification-anzsoc/2011`, [Online; accessed 21-February-2023].

O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T (2006) Uncertain judgements: eliciting experts' probabilities .

O'Hagan A (2019) Expert knowledge elicitation: subjective but scientific. *The American Statistician* 73(sup1):69–81.

Perälä T, Vanhatalo J, Chrysafi A, et al. (2020) Calibrating expert assessments using hierarchical gaussian process models. *Bayesian analysis* 15(4):1251–1280.

Press SJ (2009) *Subjective and objective Bayesian statistics: principles, models, and applications* (John Wiley & Sons).

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

Schmidt P, Witte AD (1989) Predicting criminal recidivism using 'split population'survival time models. *Journal of Econometrics* 40(1):141–159.

Thomas O, Pesonen H, Corander J (2020) Probabilistic elicitation of expert knowledge through assessment of computer simulations. *arXiv preprint arXiv:2002.10902* .

Tollenaar N, van der Heijden PG (2013) 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.

Wang C, Bier VM (2013) Expert elicitation of adversary preferences using ordinal judgments. *Operations Research* 61(2):372–385.

Winkler RL (1967) The quantification of judgment: Some methodological suggestions. *Journal of the American Statistical Association* 62(320):1105–1120.

Zyphur MJ, Oswald FL (2015) Bayesian estimation and inference: A user's guide. *Journal of Management* 41(2):390–420.