Conformal Prediction for Resource Prioritisation in Predicting Rare and Dangerous Outcomes

Varun Babbar
University of Cambridge
varundbabbar@gmail.com

Umang Bhatt
University of Cambridge
The Alan Turing Institute
usb20@cam.ac.uk

Miri Zilka
University of Cambridge
mz477@cam.ac.uk

Adrian Weller
University of Cambridge
The Alan Turing Institute
aw665@cam.ac.uk

Abstract

In a growing number of high-stakes decision-making scenarios, experts are aided by recommendations from machine learning (ML) models. However, predicting rare but dangerous outcomes can prove challenging for both humans and machines. Here we simulate a setting where ML models help law enforcement prioritise human effort in monitoring individuals undergoing radicalisation. We discuss the utility of set-valued predictions in guaranteeing the maximal rate at which dangerous radicalized individuals are missed by an assisted decision-making system. We demonstrate the trade-off between risk and the required human effort. We show that set-valued predictions can help better allocate resources whilst controlling the number of high-risk individuals missed. This work explores using conformal prediction and more general risk control methods for assisting in predicting rare and critical outcomes, and developing methods for more expert-aligned prediction sets.

1 Introduction

In high-stakes settings, the consequences of misclassification can be disastrous, especially when a high-risk instance is labeled as low risk. For these applications, we can not rely on automated decision making unless our machine learning (ML) models achieve extraordinarily high performance. As this is often not the case, in practice the decision making either remains fully human, or a ML model is used to assist the human decision maker. Well designed human-AI teams can improve both performance Bansal, Wu, and Zhou (2021), fairness Keswani, Lease, and Kenthapadi (2021), and trust Bansal et al. (2020). Researchers have explored applications in content moderation Link, Hellingrath, and Ling (2016); Jhaver et al. (2019), medical imaging, Fogliato et al. (2022); Hamid et al. (2018), and risk assessments Green and Chen (2019). In this work, we are interested in leveraging collaboration to archive better risk control. Specifically, we consider the challenge of predicting a rare and dangerous label, where mis-classification can lead to serious real-world consequences. We aim to demonstrate the potential value of incorporating a predictive model in a collaborative setting, even if the model alone cannot predict the label of interest to a satisfactory level.

There are several non-parametric, distribution free risk controlling methods found in literature Bates et al. (2020); Angelopoulos et al. (2021, 2022); Vovk, Gammerman, and Shafer (2005). These can provide actionable uncertainty quantification for any black box classification model in the form of predictive sets with theoretically guaranteed risk levels below the desired amount (e.g. false negative

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

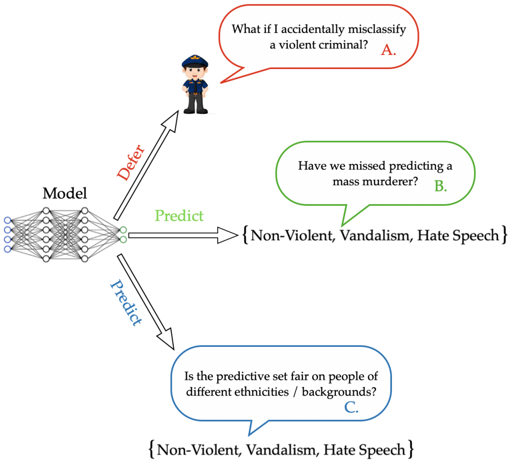

What if I accidentally misclassify a violent criminal? A.
Defer
Model
Predict
Have we missed predicting a mass murderer? B.
{Non-Violent, Vandalism, Hate Speech}
Predict
Is the predictive set fair on people of different ethnicities / backgrounds? C.
{Non-Violent, Vandalism, Hate Speech}

Figure 1: An illustration of the types of risk we may want to control:
A: Expert Misclassification Risk.
B: False Negative Rate.
C: Fairness with Respect to Different Groups
Lu et al. (2022) has focused on CP methods for ensuring fairer predictive sets, but they didn't formulate this as a risk control problem. In this paper, we focus on Risks A and B and leave controlling for risk C for future work.

rates). Angelopoulos et al. (2021) have further developed a scheme that further extends risk control to obtain predictive sets that can control for multiple risks. However, most such prior work and others such as Angelopoulos et al. (2020); Romano, Sesia, and Candes (2020); Stutz et al. (2022) has relied on the predictive quality of the model for the success of the approach and did not considered control of multiple risks in a human-AI team. While work such as Babbar, Bhatt, and Weller (2022) has considered the impact of prediction sets in human-AI teams, they a) only considered conformal prediction (CP) and b) did not consider how CP (and indeed, broader risk control methods) can be leveraged to provide guarantees on the false negative rates of the expert.

In the context of human-AI teams, we put risk controlling procedures in the context of a human expert. Concretely, we explore risk controlling predictive sets when there is a resource constraint on the downstream decision maker, i.e. they can only perform predictive duties / take further action on a select few examples. In these situations, we want satisfactory guarantees on the expert's misclassification rate (without making assumptions on the expert's competencies) and allocate the right examples to the expert for further monitoring and intervention. For the human-AI team as a whole, we may also want to control for other risks associated with the model. Some risks we may want to control are illustrated in Figure 1.

Through our exploration of a case study involving flagging dangerous radicalised individuals for further intervention, we argue that risk control is a multi-faceted problem that needs to be carefully considered from the perspective of both the downstream decision maker and the automated assistant insofar as cost sensitive classification is concerned. Despite any intrinsic limitations of the classifier and the underlying expert, this paper serves provide reassurance to the end user that any predictive sets output by a classifier will be useful, trustworthy, and faithfully articulate model uncertainty.

2 Use Case: Radicalization in the US

In this work, we use the following motivating example: identifying individuals undergoing a dangerous radicalisation process. Following McFee, Jensen, and James (2019), we consider a scenario where a law enforcement agency has limited information on a large volume of individuals, and an intervention is required to procure additional information. This intervention (e.g. monitoring chatter) requires resource and justification; therefore, it is critical to prioritise which individuals may pose a high risk to society. We use the available data to build a model that predicts the extent of radical activity the individual in question may go on to commit. We then use the model outcomes to prioritise the available resources on the most crucial interventions.

We build a predictive model on the Profiles of Individual Radicalization in the United States (PIRUS) dataset McFee, Jensen, and James (2019). This public dataset contains anonymised information on the socioeconomic background, personality, childhood, ideology, and radicalization process of 2226 individuals in the USA who have a history of extremist activities, either violent or non-violent in nature, from 1948 to 2018. The aim of this dataset is to enable the development of methods that help understand the process of radicalization from a scientifically rigorous perspective and help mitigate this process in its nascent stages. Although all the information in this dataset was collected post-hoc, we attempted to only use information that could have reasonably been available before an act of violence was committed as input features. We combined outcome information from the dataset to engineer a target label indicating the level of harm extracted by the individual.

Broadly speaking, the input features can be divided into 3 categories:

- **Group Nature:** Includes extremist group dynamics and recruitment mechanisms of the group the individual was associated with.
- **Radicalisation:** Information about the extent of radicalisation and exposure of the individual to different radicalisation material.
- **Other information:** Demographics, socioeconomic status, and personal information about social relationships and prior criminal activity.

We construct the target label from information about the first publicly known extremist activity associated with an individual. Specifically, we considered whether an individual's violent plot was executed according to their plan, and the resulting casualties. See Table 2 in the Appendix for details.

As we build a model to decide where to allocate resources to investigate potentially dangerous radicals, we consider how to best facilitate cooperation between experts and the model. In spirit of Babbar, Bhatt, and Weller (2022), we provide set-valued predictions, which capture the uncertainty associated with a model's outcome. We consider how experts can help shape the multiple risks for which we control, as we may care about metrics beyond accuracy. Specifically, we can consider learning a deferral risk, whereby the model abstains from performing prediction on individuals where the model is uncertain. We also consider how to triage the number of deferred individuals under fixed resources.

3 Preliminary Theoretical Background

We define a set valued prediction Γ_λ as a mapping from the input space $\mathcal{X}$ to the power set of the label space $\mathcal{Y}$, i.e. $\Gamma_\lambda(X) : \mathcal{X} \to 2^{\mathcal{Y}}$. In general, this would be a wrapper for a classifier $m_\theta(x)$ that provides softmax probabilities $\pi_y(x)$. We construct a risk controlling set predictor as:

$$\Gamma_\lambda(x) = \{y : \tau(x, y) \geq \lambda\} \tag{1}$$

where $\tau(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathcal{R}$ is a non-conformity score function. This is a function that quantifies how different example (X, y) is from previously observed data. In this paper, for all risk controlling sets, we use the conformity score function in Sadinle, Lei, and Wasserman (2016), which is defined as $\tau(x, y) = \pi_y(X)$, i.e. the softmax probability associated with label y and example X.

3.1 Conformal Prediction

In Conformal Prediction (CP), we tune λ to construct sets that satisfy a desired false negative tolerance, i.e:

$$P(Y_{test} \in \Gamma_\lambda(X_{test})) \geq 1 - \alpha \tag{2}$$

for any chosen α. A validation dataset is split into calibration and test datasets. The calibration dataset is used to determine the largest value of λ such that Equation 2 holds.

3.2 Generalised Risk Control

The guarantees provided by CP hold in expectation and are only valid when the desired risk to be controlled is the false negative rate. However, to express model uncertainty and provide useful sets to the downstream decision maker, we may want to control for a broader class of risks. Furthermore, we may seek to limit the downside violation of such risks wherever possible (i.e. with high probability). We now want to be able to find an appropriate threshold λ to obtain a set predictor that is guaranteed to control for potentially multiple risks simultaneously with high probability. Bates et al. (2021) developed a procedure called Risk Controlling Prediction Sets (RCPS) that can control for any risk function(s) at the user-desired level, i.e.

Theorem 3.1. *We can control for any risk $R(\lambda) = \mathbb{E}[L(Y, \Gamma_\lambda(X)]$ at a user-specified level α with probability at least $1 - \delta$ for a user-defined loss function $L(Y, \Gamma_\lambda(X))$:*

$$P(R(\lambda) \leq \alpha) \geq 1 - \delta \tag{3}$$

Proof. From Bates et al. (2020) □

Using RCPS, we can generate risk controlling procedures that aim to diversify risk between the model and the human. Concretely, we employ ideas from rejection learning found in literature Mozannar and Sontag (2020) to enable the model to reject some examples for the expert to classify. We now want to simultaneously ensure that the model does not perform badly on examples it needs to classify and the expert is reasonably good on examples it receives from the model. However, we argue that risk control should not end here - humans may want to control for a variety of other risks associated with a prediction. In this paper, we make no limiting assumptions about the kind of risk functions humans can provide. Rather, we aim to illustrate multiple risk control in cost sensitive scenarios, regardless of the definition of risk. Thus, given a human $h(X)$ and a rejector $r(X) \in \{0, 1\}$ (where $r(X) = 1$ implies deferral to $h(X)$), we can apply RCPS and individually control for multiple risks.

Corollary 3.1.1. *We can control for the misclassification rate of the human at a chosen level and any other risk function simultaneously with high probability, i.e.*

$$P(P(h(X) \neq Y | r(X) = 1) \leq \alpha_1) \geq 1 - \delta \tag{4}$$

$$P(R(\lambda) = \mathbb{E}[L(Y, \Gamma_\lambda(X)) | r(X) = 0] \leq \alpha_2) \geq 1 - \delta \tag{5}$$

for suitable α_1, α_2, δ

Proof. See Appendix. □

In this paper, we employ the following risk functions for the trained model employed on non-deferred examples:

$$\text{False Negative Rate: (FNR)} = \mathbb{E}[\mathbb{I}_{Y \notin \Gamma_\lambda(X) | Y \in \mathcal{G}}] \tag{6}$$

$$\text{False Positive Rate: (FPR)} = \mathbb{E}[\mathbb{I}_{y \in \Gamma_\lambda(X) | y \neq Y, y \in \mathcal{G}}] \tag{7}$$

where $\mathcal{G}$ is the set of predefined risky labels **(see Appendix)**. A high false negative identification rate of dangerous individuals is unquestionably disastrous. Equally, a high false positive rate can lead to wasted utilization of resources monitoring individuals that will likely not cause social harm.

4 Human-Centric Risk Control Experiments

We perform experiments on the PIRUS dataset McFee, Jensen, and James (2019) to illustrate different aspects of human-centric risk control. We aim to show the following:

- We can generate set predictors that simultaneously control for the marginal false negative identification rate of more dangerous individuals (i.e. cases where an extremist assaulted civilians with a deadly weapon) and the misclassification rate of the expert.
- We analyse the conditions under which multiple risk control is achievable. Specifically, we compare the best risk control possible under different deferral policies.
- We then employ risk controlling predictive sets in a situation where law enforcement may want to monitor a limited number of individuals and apply interventionist policies. In these scenarios, we explore a few intervention methods based on predictive sets.

Babbar, Bhatt, and Weller (2022) have recently evaluated a scheme called D-CP, where they learn a set predictor that defers some examples to an expert and and provides calibrated predictions on non-deferred examples. We extend this scheme now in order to provide statistical guarantees on the expert's misclassification rate.

We first generate 4 synthetic experts with the following characteristics:

- If the true label is not dangerous, the expert randomly assigns any label to the example.
- If the true label is dangerous, i.e. is either $5, 6$, or 7, the expert can classify the label with the following accuracies – Expert A: 95%, Expert B: 90%, Expert C: 80%, and Expert D: 70%

We now want a suitable deferral policy such that we defer a non-zero number of examples to such an expert whilst ensuring that the expert is right $1 - \gamma_1$ proportion of the time, on future examples, with high probability. Moreover, whenever the true label is dangerous and we don't defer, we also want the set to contain the label $1 - \gamma_2$ proportion of the time. To this end, given an expert $h(X)$, we define the set predictor

$$\Gamma_\lambda(X) = \begin{cases} \emptyset & r(X) \geq \lambda_1 \\ \{y : \tau(X, y) \geq \lambda_2\} & \text{otherwise} \end{cases}$$

where $r(X) \in [0, 1]$ is the rejector trained using the loss function + procedure in Mozannar and Sontag (2020) and deferral is equivalent to $\Gamma_\lambda(X) = \emptyset$. (See Appendix for training details). Next, we define the risk functions that satisfy the above aforementioned requirements:

$$R_1(\lambda_1) = P(\Gamma_\lambda(X) = \emptyset | h(X) \neq Y, Y \in \mathcal{G}) \tag{8}$$

$$R_2(\lambda_1, \lambda_2) = P(Y \notin \Gamma_\lambda(X) | \Gamma_\lambda(X) \neq \emptyset, Y \in \mathcal{G}) \tag{9}$$

The equivalent empirical risks are:

$$\hat{R}_1(\lambda_1) = \frac{1}{\sum_{j=1}^N \mathbb{I}_{r(X_j) \geq \lambda_1, Y_j \in \mathcal{G}}} \sum_{i=1}^N \mathbb{I}_{h(X_i) \neq Y_i} \mathbb{I}_{r(X_i) \geq \lambda_1, Y_i \in \mathcal{G}} \tag{10}$$

$$\hat{R}_2(\lambda_1, \lambda_2) = \frac{1}{\sum_{j=1}^N \mathbb{I}_{r(X_j) \leq \lambda_1, Y_j \in \mathcal{G}}} \sum_{i=1}^N \mathbb{I}_{Y_i \notin \Gamma_\lambda(X_i)} \mathbb{I}_{r(X_i) \leq \lambda_1, Y_i \in \mathcal{G}} \tag{11}$$

Here, λ_1 is the threshold for deferral, i.e. whenever the model's prediction for the deferral class $\pi_\perp(X) \in [0, 1] \geq \lambda_1$ we defer. λ_2 is the threshold for including any label in the set. We first tune λ_1 using the RCPS procedure to find the smallest λ_1 such that:

$$P(R_1(\lambda_1) \leq \gamma_1) \geq 1 - \delta \tag{12}$$

Then, we fix λ_1 and tune λ_2 using RCPS such that

$$P(R_2(\lambda_1, \lambda_2) \leq \gamma_2) \geq 1 - \delta \tag{13}$$

We then train the deferral policy on all 4 experts. During test time, we perform 1000 random splits of the calibration and validation dataset and determine the expert and model risk for each split. For each expert, we set the level of risk control as the corresponding misclassification rate, i.e. $\gamma_1 \in \{0.05, 0.1, 0.2, 0.3\}$. The distribution of these risks is shown in the violin plot in Figure 3. For each expert, we are able to achieve precise control of the expected misclassification rate. The probability mass above the desired risk level is very small - we found this to be smaller than δ for the values tested, demonstrating that we satisfied the guarantee in Equation 12. Note that in general, we were not able to achieve precise control for γ_1 when it is much lower than the expert's error rate (for smaller risks, the trivial solution found by the scheme is to defer no examples). This is likely due to inherent limitations of the deferral policy - because it is trained on finite, noisy data, it cannot always accurately gauge which examples an expert will be correct on. However, on the non deferred examples, we obtain precise control of the false negative rate (i.e. the probability the true label is not in the set) for labels $\mathcal{G} = [5, 6, 7]$. This guarantee is independent of the expert we defer examples to.

Takeaway: We can use the RCPS procedure to simultaneously obtain statistical guarantees on the false negative rate of the model and the accuracy of the expert. This is more likely to be useful when labels are not too rare to divide responsibilities.

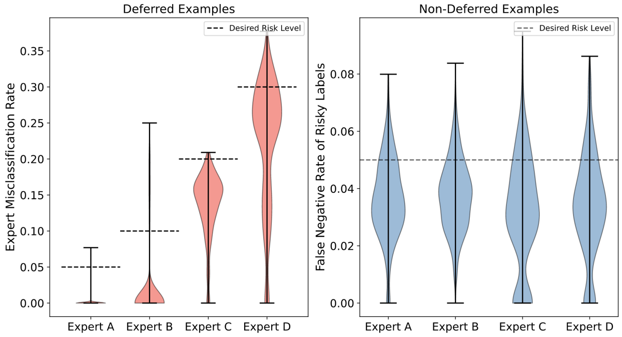


Figure 2: Illustration of dual risk control on risky labels over $N = 1000$ calibration-test dataset splits (**left**: Expert Misclassification Rate, **right**: FNR for instances where the extremist assaulted civilians with a deadly weapon). We want there to be at most $\delta = 0.1$ probability of the risk being greater than the desired risk level highlighted in the plots.

Table 1: Some Flagged and Non-Flagged Individuals

Individual ID	Feature 1	Feature 2	Feature 3	Feature 4	Predictive Set	Flag
9411	Participated in extremist dialogue	Far-right Islamist	Evidence of Mental Illness	Previously committed violent crimes	[6,7]	Yes
1669	Primarily radicalised via the internet	White supremacist	No Evidence of Mental Illness	Had a friend who committed violent crimes	[6,7]	Yes
4150	Had friends who influenced radicalisation	Environmentalist / Animal Rights Activist	No Evidence of Mental Illness	Never committed any crime	[0,1,2]	No

5 Recommending Dangerous Individuals for Profiling

We now focus on risk control purely from the perspective of the model - we want the model to be able to recommend and prioritise intervention for individuals who are likely to carry out large scale attacks in the future. In this case, they are grouped under the most dangerous label: 7. As these individuals will be rare, we now ask the question: To what extent can a risk-control scheme provide useful information for intervention as the rarity of the dangerous label increases? It can be impractical to provide simultaneous risk control for the expert and the model in these situations as the labels are too rare to divide predictive responsibilities.

Instead, we now assume that the human expert has a budget β which they can expend on intervention, profiling, and monitoring of potentially dangerous individuals. The unit cost of an intervention is C. A model acting within the expert's budget parameters needs to flag a limited number of individuals for monitoring whilst ensuring that the false negative rate is acceptably low. This is different from deferring examples as the model is now actively providing set valued predictions on each example. To investigate this, we use a generative model $\mathcal{G} : \mathcal{L} \to \mathcal{X}$ which takes Gaussian noise from the latent space $\mathcal{L}$ as input and generates a new data point in the form of a feature matrix that arises from the same distribution as all high-risk individuals seen in the training data for the model. We generate increasing numbers of such synthetic datapoints, add them to the existing training set of all examples, and train our set valued predictor on the new training set. We now generate predictive sets that control the FNR for label 7 with $\gamma_2 = 0.05$ and $\delta = 0.1$. In this case, the risk function is $R(\lambda) = P(Y \notin \Gamma_\lambda(X) | Y = 7)$. Then, we choose the $\lfloor \frac{\beta}{C} \rfloor$ examples for intervention in the following manner:

- **Smallest set:** We choose the $\lfloor \frac{\beta}{C} \rfloor$ examples with the smallest predictive set size. Of these, we intervene when the set contains either $5, 6$, or 7.
- **Sets with the highest average label**
- **Most probable label:** We intervene when the most probable label is $5, 6$ or 7.
- **Random Allocation of Labels**

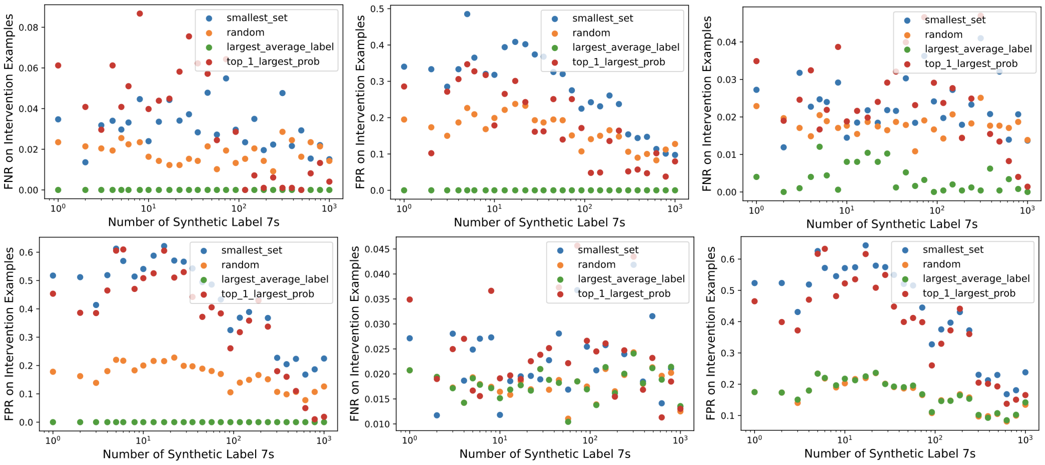


Figure 3: (Left) Budget $\beta = 50$, cost $C = 1$. (Middle) Budget $\beta = 250$, cost $C = 1$. (Right) Budget $\beta = 1000$, cost $C = 1$. Risk controlling predictive sets were constructed with risk $\gamma = 0.05$ and $\delta = 0.1$

For each strategy, we measure the FNR and FPR, defined in Equations 6 and 7 respectively. Because we are employing ordinal risk control, we are assured of small predictive sets that contain a dangerous label with high probability. These examples can then be shown to the end user alongside the predictive set and an indication that they should be flagged for further review. Some examples of flagged individuals are shown in Table 1.

From Figure 3, we note the following:

- For lower budgets, compared to other methods, selecting examples with predictive sets containing high average labels provides the best FPR and FNR with regards to intervention / monitoring of dangerous individuals.
- However, there is not much difference between random allocation and high average label allocation when we intervene on a large number of examples (i.e. with a high budget). This makes sense because after allocating individuals most likely to be dangerous, there are fewer such individuals in the remaining test pool, leading to higher FPRs and FNRs.

6 Conclusion and Future Work

In this study, we explored an application of risk controlling methods in predicting whether radicalised individuals will go on to commit violent acts. We explore risk control for Human-AI teams by considering two scenarios:

- The model defers some examples to an underlying human expert, for example, law enforcement. In this situation, we apply risk controlling procedures developed by Angelopoulos et al. (2021) and Bates et al. (2020) to simultaneously control for the expert misclassification rate and the model's misclassification rate of dangerous individuals. The level of controllable risk of the expert depends on the underlying deferral policy and the number of examples deferred. Future work could explore other deferral policies such as those in Okati, De, and Gomez-Rodriguez (2021)
- In an alternate scenario, the model doesn't abstain from predictions. However, we enforce a resource constraint wherein only a limited number of individuals can be further monitored, profiled, and interventionist policies applied when necessary. For this situation, we explore 4 different heuristics for allocating the examples based on their predictive set size. We find that intervening on examples with the highest average label in the predictive set outperforms other baselines such as random allocation, most probable label allocation, and smallest set based allocation, i.e. it achieves a lower false positive and false negative rate of dangerous individuals on allocated examples. We leave exploration of more rigorous allocation methods for future work.

Ultimately, human-AI collaboration in all its forms is of utmost importance in safety-critical domains like criminal justice. Providing predictive sets as a risk control strategy allows the downstream decision maker to better interpret the uncertainty associated with the model. We hope this paper promotes future work on better allocation policies for intervention, control of other risk functions that humans may find useful, and other applications where risk control can be used to quantify and mitigate high risk events.

Acknowledgements

UB acknowledges support from DeepMind and the Leverhulme Trust via the Leverhulme Centre for the Future of Intelligence (CFI), and from the Mozilla Foundation. MZ acknowledges support from the Leverhulme Trust grant ECF-2021-429. AW acknowledges support from a Turing AI Fellowship under grant EP/V025279/1, The Alan Turing Institute, and the Leverhulme Trust via CFI.

References

Angelopoulos, A. N.; Bates, S.; Jordan, M.; and Malik, J. 2020. Uncertainty Sets for Image Classifiers Using Conformal Prediction. In *International Conference on Learning Representations*.

Angelopoulos, A. N.; Bates, S.; Candès, E. J.; Jordan, M. I.; and Lei, L. 2021. Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control. *CoRR* abs/2110.01052.

Angelopoulos, A. N.; Bates, S.; Fisch, A.; Lei, L.; and Schuster, T. 2022. Conformal risk control.

Babbar, V.; Bhatt, U.; and Weller, A. 2022. On the Utility of Prediction Sets in Human-AI Teams. *ArXiv* abs/2205.01411.

Bansal, G.; Nushi, B.; Kamar, E.; Horvitz, E.; and Weld, D. S. 2020. Is the most accurate ai the best teammate? optimizing ai for teamwork.

Bansal, G.; Wu, T.; and Zhou, J. 2021. Does the whole exceed its parts? the efect of ai explanations on complementary team performance. Association for Computing Machinery.

Bates, S.; Angelopoulos, A. N.; Lei, L.; Malik, J.; and Jordan, M. I. 2020. Distribution-free, risk-controlling prediction sets. *arXiv preprint arXiv:2101.02703*.

Bates, S.; Angelopoulos, A.; Lei, L.; Malik, J.; and Jordan, M. 2021. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)* 68(6):1–34.

Fogliato, R.; Chappidi, S.; Lungren, M.; Fitzke, M.; Parkinson, M.; Wilson, D.; Fisher, P.; Horvitz, E.; Inkpen, K.; and Nushi, B. 2022. Who goes first? influences of human-ai workflow on decision making in clinical imaging.

Green, B., and Chen, Y. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 90–99. New York, NY, USA: Association for Computing Machinery.

Hamid, K.; Asif, A.; Abbasi, W.; Sabih, D.; and Minhas, F. 2018. Machine learning with abstention for automated liver disease diagnosis.

Jhaver, S.; Birman, I.; Gilbert, E.; and Bruckman, A. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Trans. Comput.-Hum. Interact.* 26(5).

Keswani, V.; Lease, M.; and Kenthapadi, K. 2021. Towards unbiased and accurate deferral to multiple experts. *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* 154–165.

Link, D.; Hellingrath, B.; and Ling, J. 2016. A human-is-the-loop approach for semi-automated content moderation. In *ISCRAM*.

Lu, C.; Lemay, A.; Chang, K.; Höbel, K.; and Kalpathy-Cramer, J. 2022. Fair conformal predictors for applications in medical imaging. *Proceedings of the AAAI Conference on Artificial Intelligence* 36(11):12008–12016.

Madras, D.; Pitassi, T.; and Zemel, R. 2018. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 6150–6160.

McFee, G.; Jensen, M.; and James, P. 2019. Profiles of individual radicalization in the united states (pirus). *College Park, MD: National Consortium for Terrorism and Responses to Terrorism, University of Maryland. Retrieved September* 17:2019.

Mozannar, H., and Sontag, D. 2020. Consistent Estimators for Learning to Defer to an Expert. In *International Conference on Machine Learning*, 7076–7087. PMLR.

Okati, N.; De, A.; and Gomez-Rodriguez, M. 2021. Differentiable Learning Under Triage. In *Advances in Neural Information Processing Systems*.

Romano, Y.; Sesia, M.; and Candes, E. 2020. Classification with Valid and Adaptive Coverage. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 3581–3591. Curran Associates, Inc.

Sadinle, M.; Lei, J.; and Wasserman, L. 2016. Least Ambiguous Set-Valued Classifiers with Bounded Error Levels. *Journal of the American Statistical Association* 114:223–234.

Stutz, D.; Dvijotham, K. D.; Cemgil, A. T.; and Doucet, A. 2022. Learning optimal conformal classifiers. In *ICLR*.

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic Learning in a Random World*. Springer.

Wilder, B.; Horvitz, E.; and Kamar, E. 2021. Learning to Complement Humans. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

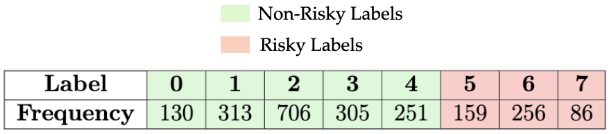
A Pre-Processing of the PIRUS Dataset

We observed the following characteristics of the PIRUS dataset

- The dataset contains noisy entries with missing features. To deal with this issue, we replaced all instances of missing feature j with the average feature value found across the dataset.
- The dataset is unlabelled, prompting the need to generate synthetic labels for further analysis. We do so using the procedure outlined in Table 2. As the labels are generated using select features, in the event that one of the relevant features is missing for an instance, we remove that instance from the dataset. This was done for 20 out of 2226 datapoints.
- For each instance, the dataset contains timestamps reflecting various events such as the approximate date of exposure to radical ideologies, and date of religious conversion (if applicable). We converted dates to UNIX timestamps - which represent seconds passed since 00:00:00 UTC, Jan 1970.
- The dataset contains features of different scales. To counteract this, we performed feature normalization, i.e. for the j^{th} feature of the i^{th} instance, we replaced the feature value by:

$$\hat{x}_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j} \qquad \hat{\mu}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij} \qquad \hat{\sigma}_j^2 = \frac{1}{N}\sum_{i=1}^{N}(x_{ij} - \mu_j)^2 \tag{14}$$

  for all features.



🟩 Non-Risky Labels

🟥 Risky Labels

Label	0	1	2	3	4	5	6	7
Frequency	130	313	706	305	251	159	256	86

Figure 4: PIRUS Dataset Label Distribution

Table 2: The criterion used for generating labels from 3 features found in the PIRUS dataset. We place high emphasis on severity of crimes already committed as we consider them to be a meaningful indicator of future risk of violence. However, future research could consider alternate methods of label generation

Label	Plot Extent (1-5)	Criminal Severity (0-10)	Anticipated Fatalities
7	5	10	Greater than 20
6	5	10	Greater than 1
5	5	10	None
4	< 5	10	Any
3	Any	8/9	Any
2	Any	5-7	Any
1	Any	3/4	Any
0	Any	0-2	Any

After pre-processing and feature generation, we trained a simple Multilayer Perceptron on Google Colab with learning rate $\eta = 0.005$ on the dataset for 30 epochs. We used the loss function in Mozannar and Sontag (2020) to train the classifier over an augmented label space with deferral as an additional class.

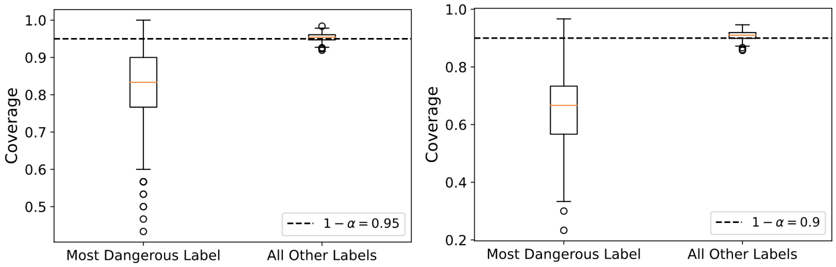


Figure 5: Class conditional coverage on the PIRUS dataset provided by marginal CP using the method in Sadinle, Lei, and Wasserman (2016) (left: $1 - \alpha = 0.95$, right: $1 - \alpha = 0.9$). CP undercovers the most dangerous label relative to the desired level as coverage is only guaranteed in expectation. A better alternative is to provide rigorous, more general risk control guarantees on dangerous labels.

B Proofs

Corollary B.0.1. *We can control for the misclassification rate of the human at any desired level and any other risk function simultaneously with high probability, i.e.*

$$P(P(h(X) \neq Y | r(X) = 1) \leq \alpha_1) \geq 1 - \delta \tag{15}$$

$$P(R(\lambda) = \mathbb{E}[L(Y, \Gamma_\lambda(X)) | r(X) = 0] \leq \alpha_2) \geq 1 - \delta \tag{16}$$

for any α_1, α_2, δ

Proof. We draw inspiration from the Learn Then Test (LTT) procedure in Angelopoulos et al. (2021) for Out-Of-Distribution (OOD) detection. This is a generalisation of RCPS in that it is designed to control for risks that are not necessarily monotonic with respect to the threshold parameter λ. Here, the equivalent OOD example would be one the expert is correct on, and we would defer this example. Thus, we mark an example where the model defers as OOD. We want to defer some examples while controlling for the risk of the model deferring an example the expert makes a mistake on. This is equivalent to the risk of marking an example as OOD when it is actually in distribution, i.e. a false positive in a sense. Define the risk functions

$$R_1(\lambda_1) = P(\Gamma_\lambda(X) = \emptyset | h(X) \neq Y) \tag{17}$$

$$R_2(\lambda_1, \lambda_2) = P(Y \notin \Gamma_\lambda(X) | \Gamma_\lambda(X) \neq \emptyset) \tag{18}$$

where deferral is equivalent to outputting an empty set $\emptyset$. λ_1 is the threshold for deferral, i.e. whenever $r(X) \in [0, 1] \geq \lambda_1$ we defer and λ_2 is the threshold for including any label in the set (e.g. but not limited to the threshold conformity score τ_{cal}). That is:

$$\Gamma_\lambda(X) = \begin{cases} \emptyset & r(X) \geq \lambda_1 \\ \{y : \tau(X, y) \geq \lambda_2\} & \text{otherwise} \end{cases}$$

As above, we calculate UCBs R_1^+ and R_2^+ for all λ_1 and λ_2. Then, we choose a $\hat{\lambda}_1 \in \hat{\Lambda}_1$ where $\hat{\Lambda}_1 = \{\lambda \in \Lambda_1 : R_1^+(\lambda') \leq \gamma_1,\ \forall \lambda' > \lambda\}$ and a $\hat{\lambda}_2 \in \hat{\Lambda}_2$ where $\hat{\Lambda}_2 = \{\lambda \in \Lambda_2 : R_2^+(\hat{\lambda}_1, \lambda') \leq \gamma_2,\ \forall \lambda' > \lambda\}$. The guarantees in Equations 15 and 16 follow hence. Note that not all values of γ_1, γ_2, and δ are controllable in the sense of providing non trivial deferral mechanisms - these may depend on the performance of the human / deferral policy as well as the sample size provided. For controllable risks, we may choose a particular λ depending on the deferral budget. For example, if we want to defer 20% of examples, we may choose an appropriate $\lambda \in \Lambda'$ and obtain the same guarantees. □

C Risk Control with a Deferral Constraint

While dual risk control is attractive, Angelopoulos et al. (2021) highlight that not all risks can be controlled at the desired level. This may be, for example, due to finite calibration sample size, inherent limitations of the expert, suboptimality of the deferral policy, etc. In this section, we investigate how risk control changes as we change deferral policies and expert types. In particular, we first generate a synthetic expert which has access to some side features not used in training. These include data about prior criminal charges, the extent to which any previous attacks were planned, and degree of preparedness of the criminal who conducted an attack. The synthetic expert with access to these features has $\approx 89\%$ accuracy on risky deferred test examples. We then train 3 deferral policies using the procedure in Mozannar and Sontag (2020). Note that we can also use other methods found in literature, e.g. Wilder, Horvitz, and Kamar (2021); Madras, Pitassi, and Zemel (2018); Okati, De, and Gomez-Rodriguez (2021).

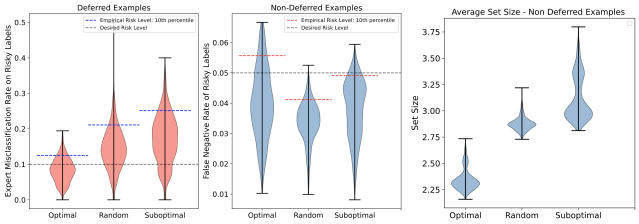


Figure 6: Comparison between deferral policies and their corresponding empirical risk control levels for a deferral rate of $\beta \leq 0.1$, $\delta = 0.1$ (**left**: expert risk $\hat{R}_1(\lambda_1)$ - desired level $\gamma_1 = 0.1$, **right**: model risk $\hat{R}_2(\lambda_1, \lambda_2)$ - desired level $\gamma_2 = 0.05$, **middle**: Distribution of average set sizes of non-deferred examples over 1000 trials). The controllability of expert misclassification risk and the resulting set size on non-deferred examples is inherently decided by the type of deferral policy trained.

- **Optimal**: The policy is trained on synthetically generated expert labels.
- **Random**: This is a deferral policy that doesn't learn the expert's strengths and defers at random. It is trained using random labels.
- **Suboptimal**: Here, we use the deferral policy that is trained to defer whenever the expert errs, i.e. we try and defer whenever the policy thinks the expert is wrong. This is done by providing inverted expert labels to the policy.

With the set predictor in Equation C, we now control for the deferral rate β in the following manner: We first generate a set of λ_1's that ensure that the deferral rate is less than β. Out of these, we choose the λ_1 such that the $1 - \delta$ UCB of $\hat{R}_1(\lambda_1)$ is closest to γ_1. This procedure is performed over 1000 random splits of the calibration and validation sets.

From Figure 6, we see that the optimal deferral policy controls for the lowest expert misclassification rate and is closest to the desired level ($\gamma_1 = 0.1$). This serves to illustrate that the degree of expert misclassification risk control is also inherently decided by the optimality of the deferral policy. Furthermore, risk control on non-deferred examples is independent of the deferral policy used - this is because dual risk control essentially combines two independent parameterised risks. Note that we could also perform a Bayes optimal allocation of examples:

$$\Gamma_\lambda(X) = \begin{cases} \emptyset & r(X) \geq \max_y \pi_y(X) \\ \{y : \tau(X, y) \geq \lambda_2\} & \text{otherwise} \end{cases}$$

In this case, we might obtain a better expert accuracy (and consequently better risk control) than determining an appropriate threshold, but we lose fine grained control of the deferral rate - this might be a constraint in real world problems.

Takeaway: The controllability of expert misclassification risk and the resulting set size on non-deferred examples inherently depend on the type of deferral policy trained.