American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

# Uncertainty-Aware Neural Networks in High-Stakes Administrative Decisions

**Dr. Ethan R. Walker**

Assistant Professor, University of Toronto, Canada

**Abstract**

Neural networks are increasingly capable of supporting administrative decisions involving eligibility assessment, risk screening, case prioritization, fraud detection, licensing, inspections, migration processing, and allocation of public services. Conventional predictive systems, however, typically produce a class label or probability without reliably distinguishing between predictions supported by familiar evidence and predictions arising from ambiguous, incomplete, noisy, or distributionally unfamiliar cases. This limitation is particularly important in public administration because an incorrectly confident prediction can influence legal rights, access to essential services, financial interests, or other consequential outcomes. The present study develops an uncertainty-aware neural decision framework in which predictive confidence is treated as a governance signal rather than merely a numerical model output. The framework integrates probability calibration, Monte Carlo dropout, deep ensembles, conformal prediction, and selective deferral to distinguish routine cases from cases requiring additional human review. Research on modern neural networks has established that classification accuracy and probability calibration are different properties, that deep networks can become overconfident, and that uncertainty estimation may deteriorate under dataset shift. Deep ensembles have demonstrated strong predictive uncertainty performance, while conformal methods provide a principled approach to constructing prediction sets under stated statistical assumptions.

A synthetic administrative dataset containing 14,000 records was constructed, comprising 10,000 model-development records, 2,000 calibration records, and 2,000 held-out evaluation records. Four decision architectures were compared: a deterministic softmax neural network, a Monte Carlo dropout network with probability calibration, a calibrated deep ensemble, and a deep-ensemble architecture augmented with conformal prediction and selective deferral. The simulated unsafe high-confidence error rate declined from 12.4% for the deterministic architecture to 7.6%, 4.4%, and 2.1% respectively. The most uncertainty-aware architecture automatically resolved 77.6% of evaluation cases while deferring 22.4% for additional review and achieved a simulated 94.1% accuracy among automatically resolved cases. These values are methodological illustrations rather than observations from a deployed administrative system. The study argues that uncertainty-aware architectures are particularly appropriate for high-stakes public administration when uncertainty is connected to predefined escalation rules, human review, subgroup monitoring, dataset-shift detection, documentation, and effective administrative remedies. The central contribution is an operational principle of uncertainty-governed automation: an administrative AI system should not only predict an outcome but should also provide a credible indication of when its own prediction is insufficiently reliable for autonomous administrative use.

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

***Keywords:*** uncertainty-aware neural networks, administrative decision-making, predictive uncertainty, deep ensembles, conformal prediction, selective classification, probability calibration, algorithmic governance, public administration, human oversight

## 1. Introduction

Artificial intelligence is increasingly incorporated into administrative processes that influence access to public benefits, government services, educational opportunities, employment-related programs, migration procedures, regulatory inspections, taxation, licensing, and other consequential decisions. Automated systems may either make a decision directly or support a public official by producing eligibility classifications, risk scores, anomaly indicators, recommendations, or case priorities. Canada's current Directive on Automated Decision-Making explicitly recognizes that automated systems can fall within public-sector governance requirements even where they merely assist an officer and a human officially retains the final decision. The Directive is designed around administrative-law concerns such as transparency, accountability, legality, procedural fairness, testing, monitoring, and impact assessment. These developments demonstrate that the public-law significance of artificial intelligence depends not simply on whether automation is technically complete but on whether computational outputs materially shape the exercise of governmental power.

A central technical difficulty is that conventional neural networks are not inherently designed to communicate how uncertain they are about an individual prediction. Modern deep models can produce highly confident probabilities even when predictions are incorrect, and probability outputs therefore cannot automatically be interpreted as calibrated estimates of reliability. Guo, Pleiss, Sun, and Weinberger demonstrated that contemporary neural networks may be substantially miscalibrated and showed that temperature scaling can provide an effective post-hoc calibration technique in many settings. The problem becomes more serious under distribution shift. Ovadia and colleagues systematically evaluated predictive uncertainty when test conditions diverged from training conditions and found substantial deterioration across uncertainty-estimation methods, although deep ensembles performed particularly strongly among the methods considered. In administrative environments, distribution shift can arise through changing economic conditions, new applicant populations, policy amendments, data-collection changes, exceptional emergencies, or the introduction of cases that were poorly represented during model development.

Uncertainty-aware machine learning seeks to make this limitation explicit rather than concealing it behind a single score. Kendall and Gal distinguish aleatoric uncertainty, which arises from noise or ambiguity inherent in observations, from epistemic uncertainty, which reflects insufficient model knowledge and may increase when the model encounters unfamiliar inputs. Gal and Ghahramani demonstrated how dropout can be interpreted as an approximate Bayesian inference mechanism, thereby providing one practical route for estimating predictive uncertainty. Lakshminarayanan, Pritzel, and Blundell later showed that ensembles of independently trained neural networks can provide a simple and scalable approach to predictive uncertainty estimation. These approaches are particularly important in administrative AI because uncertainty can be connected to operational safeguards: the system can proceed when uncertainty remains within a predefined tolerance and defer the case when the evidence is ambiguous or unfamiliar.

A second technical development concerns selective prediction. Rather than forcing the model to issue a definitive decision for every case, selective systems are permitted to abstain when confidence is inadequate. Geifman and El-Yaniv developed SelectiveNet as a neural architecture with an integrated reject option, while subsequent research has continued to examine the accuracy–coverage trade-off created when uncertain predictions are withheld. Conformal prediction provides a complementary approach by returning prediction sets or intervals rather than an unjustifiably precise single answer. Recent implementations such as TorchCP demonstrate that conformal prediction can be applied across increasingly sophisticated machine-learning architectures, although its formal coverage properties depend on the statistical conditions under which the method is used. For administrative decisions, these methods create the possibility of distinguishing cases for which automated assistance may be reasonably dependable from those where human investigation, additional documents, or alternative procedures are necessary.

The significance of uncertainty-aware design is reinforced by contemporary AI governance frameworks. NIST's AI Risk Management Framework treats risk management as a continuous process of governing, mapping, measuring, and managing AI risks and recommends that organizations determine risk tolerances, monitor system performance, document relevant limitations, and adjust controls as circumstances change. The current consolidated EU AI Act requires appropriate levels of accuracy, robustness, and cybersecurity for high-risk AI systems and establishes requirements concerning human oversight and lifecycle risk management. Following the July 2026 AI Omnibus, Annex III high-risk-system rules are scheduled to apply from December 2, 2027, while high-risk systems embedded in covered products have an extended application date of August 2, 2028. Against this technical and regulatory background, the present study asks whether administrative AI should move from prediction-centered automation toward uncertainty-governed automation, in which uncertainty explicitly determines when a model is permitted to recommend automatic resolution and when it must defer to additional administrative review.

## 2. Objectives of the Study

### 2.1 Evaluation of Predictive Uncertainty in Administrative Neural Networks

The first objective is to examine whether neural decision systems that explicitly model predictive uncertainty can provide more reliable confidence information than a conventional deterministic neural classifier. The study compares probability calibration, high-confidence error behavior, automated decision coverage, and resolved-case accuracy across progressively uncertainty-aware architectures.

### 2.2 Assessment of Selective Deferral for High-Stakes Decisions

The second objective is to evaluate the methodological value of allowing an administrative model to abstain from making or strongly recommending a decision when uncertainty exceeds a predefined threshold. Rather than treating deferral as algorithmic failure, the study examines it as a deliberate safety mechanism that reallocates ambiguous or unfamiliar cases to human review.

### 2.3 Development of an Uncertainty-Governed Administrative AI Framework

The third objective is to develop a governance model linking predictive uncertainty with calibration monitoring, dataset-shift detection, human review, documentation, subgroup analysis, and administrative contestability. The intention is to provide a framework that can subsequently be validated using real administrative datasets under appropriate legal, ethical, and privacy safeguards.

## 3. Review of Literature

### 3.1 Predictive Uncertainty, Calibration, and Neural-Network Overconfidence

A neural classifier normally converts internal model outputs into probability-like scores through a softmax function. These scores are frequently interpreted as confidence, yet predictive confidence and empirical correctness may diverge substantially. Guo and colleagues demonstrated that modern neural networks can be poorly calibrated even when classification accuracy is high and introduced temperature scaling as a simple post-training correction capable of substantially improving calibration on many benchmark datasets. Later work has continued to refine calibration procedures, including non-parametric and extended temperature-scaling approaches. This distinction is crucial for administrative systems because a model that labels a prediction as 95% confident should not systematically be correct only 70% or 75% of the time among similarly scored cases.

Uncertainty itself is multidimensional. Kendall and Gal's distinction between aleatoric and epistemic uncertainty provides a useful conceptual framework. Aleatoric uncertainty reflects irreducible ambiguity or noise in observed information, whereas epistemic uncertainty reflects uncertainty about the model's parameters or knowledge and can potentially be reduced through additional representative data. In an administrative context, aleatoric uncertainty might arise when submitted evidence genuinely supports multiple plausible interpretations, whereas epistemic uncertainty may increase when a new applicant profile is poorly represented in training records. The distinction is operationally important because different uncertainty sources require different responses. A noisy document might require additional evidence from the applicant, whereas unfamiliar population characteristics may require model retraining, broader data collection, or temporary restriction of automated use.

Deep ensembles provide another influential approach. Lakshminarayanan and colleagues independently trained multiple neural networks and aggregated their predictive distributions, demonstrating a comparatively simple method for predictive uncertainty estimation. Ovadia and colleagues subsequently found that deep ensembles performed strongly when predictive uncertainty was tested under dataset shift, although their results also showed that uncertainty methods generally deteriorate as distribution shift becomes more severe. More recent research continues to investigate out-of-distribution uncertainty using ensemble and Bayesian approaches, reinforcing the importance of uncertainty estimation when deployment data no longer resemble model-development data.

### 3.2 Conformal Prediction and Selective Administrative Decision-Making

Selective classification modifies the conventional assumption that a model must make a prediction on every observation. Instead, the model predicts only when its uncertainty is sufficiently low and refers remaining cases to another process. Geifman and El-Yaniv's SelectiveNet formalized this principle through a neural architecture containing a reject option. Gangrade, Kag, and Saligrama similarly

describe selective classification as a framework that permits abstention and thereby trades prediction coverage against reliability. More recent work continues to evaluate post-hoc selective-classification methods designed to identify low-confidence cases without completely retraining the underlying classifier.

This framework has direct relevance to high-stakes administrative processes. A public authority may be better served by a model that automatically handles 75% of routine cases with high reliability and refers 25% for careful review than by a system that produces a definitive recommendation for 100% of cases while concealing substantial uncertainty. Deferral therefore should not automatically be interpreted as inefficiency. Its value depends on whether uncertain cases are meaningfully redirected to a process that can consider additional context rather than simply reproducing the initial algorithmic judgment.

Conformal prediction offers a complementary method for making uncertainty operational. Instead of always returning one predicted category, conformal methods may produce sets of plausible labels or intervals calibrated to a target coverage level under appropriate assumptions. Modern implementations emphasize that conformal prediction can wrap around black-box models, including neural networks, to produce distribution-free predictive sets under exchangeability or related conditions. Einbinder and colleagues specifically developed conformalized approaches for training uncertainty-aware classifiers, noting that ordinary deep neural networks are not inherently designed to provide reliable probability estimates.

### 3.3 Administrative Governance of Uncertain Predictions

Public-sector governance increasingly recognizes that algorithmic systems should be evaluated according to their impact on administrative decision-making rather than only according to whether they fully automate an outcome. Canada's Directive on Automated Decision-Making applies to systems that make or support administrative decisions and expressly includes partial automation and human-in-the-loop arrangements. Its mandatory Algorithmic Impact Assessment uses questions concerning system design, decision type, algorithmic impact, data, and mitigation measures to determine the impact level associated with an automated decision system. This approach supports the argument that uncertainty controls should be calibrated to the consequences of error rather than adopted uniformly across every administrative application.

The United Kingdom's Algorithmic Transparency Recording Standard provides another complementary mechanism. Current guidance requires covered central-government bodies and certain arm's-length organizations to publish information concerning algorithmic tools that significantly influence decisions with public effect. The standard seeks meaningful transparency not merely through acknowledging that an algorithm exists but by explaining how and why it is used. Uncertainty metrics, deferral rules, model monitoring, and the consequences of low-confidence predictions can therefore be understood as important components of meaningful algorithmic transparency.

The Council of Europe Framework Convention adopts a broader human-rights and rule-of-law perspective. It requires context-sensitive transparency and oversight and calls for iterative risk and impact assessment addressing potential consequences for human rights, democracy, and the rule of law. Its HUDERIA methodology further emphasizes variables such as scale, scope, reversibility, and

likelihood when assessing AI-related harms. An uncertainty-aware administrative architecture is compatible with this risk-based logic because it allows the acceptable level of automatic decision-making to be adjusted according to the severity and reversibility of potential harm.

## 4. Research Methodology

### 4.1 Synthetic Administrative Dataset and Experimental Design

The study uses a simulation-based comparative computational design. No actual administrative records were accessed. A fully synthetic corpus of 14,000 hypothetical administrative applications was created to represent a generalized public-benefit eligibility task. The simulation incorporated variables representing household-resource bands, employment status, application completeness, documentary consistency, declared dependants, prior program participation, income stability, and procedural indicators. Sensitive demographic attributes were not used as decision features but were simulated separately for subgroup calibration auditing.

The dataset was divided into 10,000 records for model development, 2,000 records for validation and probability calibration, and 2,000 records for final evaluation. The held-out evaluation sample intentionally contained a mixture of ordinary in-distribution cases, records with additional missing or noisy information, and simulated distribution-shift cases representing applicant patterns underrepresented during model development.Each architecture was evaluated on the same 2,000 held-out cases so that differences reflected the decision architecture rather than changes in case composition.

**Table 1: Experimental Architecture for the Synthetic Administrative Decision Study**

| Model Condition | Uncertainty Mechanism | Calibration/Decision Rule | Administrative Handling of Uncertain Cases |
|---|---|---|---|
| Deterministic Neural Network | Raw softmax confidence | No explicit post-hoc uncertainty safeguard | All cases receive an automated model outcome |
| MC Dropout + Calibration | Multiple stochastic inference passes | Temperature-calibrated probabilities; uncertainty threshold | High-uncertainty cases deferred |
| Deep Ensemble + Calibration | Five independently trained neural networks | Ensemble probability + calibration threshold | Cases with substantial model disagreement deferred |
| Deep Ensemble + Conformal Selective Deferral | Ensemble uncertainty plus conformal prediction sets | Target 90% marginal prediction-set coverage on calibration data; non-singleton or low-confidence cases deferred | Uncertain cases routed to structured human review |

**Note:** Architecture choices and all associated data are simulated for methodological demonstration.

### 4.2 Neural Architectures and Uncertainty Measurement

The deterministic baseline used a multilayer feed-forward neural classifier with a softmax output. Its highest predicted class probability was treated as conventional model confidence. This architecture represents the common practice of converting model logits into probabilities without explicitly modeling epistemic uncertainty.

The Monte Carlo dropout condition retained dropout during inference and generated 50 stochastic forward passes for each evaluation case. Mean predictive probabilities were calculated across the passes, while variation among predictions provided an approximate uncertainty signal following the Bayesian-dropout interpretation developed by Gal and Ghahramani. Temperature scaling was applied using the separate calibration dataset in accordance with the calibration principle demonstrated by Guo and colleagues.

The deep-ensemble architecture consisted of five independently initialized neural networks trained on the same synthetic development distribution. Final class probabilities were obtained by averaging the component predictions. Disagreement among ensemble members was retained as an epistemic-uncertainty indicator. The choice reflects the established deep-ensemble approach to uncertainty estimation and evidence showing comparatively strong behavior under dataset shift.

The fourth architecture combined the calibrated deep ensemble with split conformal classification. A separate calibration sample was used to construct prediction sets at a target marginal coverage level of 90%. A singleton set permitted automatic processing where the other governance thresholds were also satisfied. A prediction set containing multiple plausible administrative outcomes triggered deferral. The methodology deliberately describes the coverage target as marginal and conditional on the assumptions supporting the conformal procedure, rather than presenting it as a guarantee of individual correctness.

### 4.3 Evaluation Measures and Governance Thresholds

The first performance measure was resolved-case accuracy, defined as the proportion of correct predictions among cases that the architecture permitted to proceed without uncertainty-based deferral.

This metric differs from conventional overall accuracy because uncertainty-aware architectures intentionally decline to resolve some cases automatically.

The second measure was Expected Calibration Error (ECE), which summarizes discrepancies between predicted confidence and observed accuracy across confidence bins. Lower ECE values indicate that reported confidence more closely reflects empirical correctness.

The third measure was the unsafe high-confidence error rate, defined for this simulation as the proportion of all evaluation cases in which an incorrect model outcome was produced with calibrated confidence of at least 0.85 and was not intercepted by an uncertainty-based deferral rule.

The fourth measure was deferral rate, representing the proportion of cases routed to structured human review because the predictive uncertainty exceeded the decision threshold or because the conformal prediction set did not contain a unique admissible outcome.

No inferential p-values are reported because the data were deliberately generated rather than sampled from a real administrative population. Statistical significance on synthetic records could produce an unjustified appearance of empirical validation.

## 5. Analysis

### 5.1 Calibration and Reliability of Automated Resolution

The deterministic neural network automatically processed all 2,000 evaluation cases because it contained no reject mechanism. Its simulated resolved-case accuracy was 80.8%, with an Expected Calibration Error of 0.142. The result illustrates the difference between producing probabilities and producing reliable probabilities: the baseline architecture frequently assigned strong confidence to predictions whose empirical correctness did not justify the reported certainty.

The MC dropout architecture improved calibration to an ECE of 0.081. Its automatic coverage declined to 91.2% because 8.8% of cases were deferred under the uncertainty threshold, while accuracy among automatically resolved cases increased to 85.0%.

The calibrated deep ensemble generated a larger change. Its ECE declined to 0.044, and automatic decision coverage fell to 85.8%. Among the cases that remained eligible for automated resolution, simulated accuracy reached 89.6%.

The ensemble-plus-conformal architecture achieved the lowest calibration error at 0.028. It automatically resolved 77.6% of cases and deferred 22.4%. Accuracy among the automatically resolved group increased to 94.1%.

**Table 2: Simulated Performance of Uncertainty-Aware Neural Decision Architectures**

| Neural Decision Architecture | Automatic Decision Coverage | Deferral Rate | Resolved-Case Accuracy | Expected Calibration Error | Unsafe High-Confidence Error Rate |
|---|---|---|---|---|---|
| Deterministic neural network | 100.0% | 0.0% | 80.8% | 0.142 | 12.4% |
| MC dropout + calibration | 91.2% | 8.8% | 85.0% | 0.081 | 7.6% |
| Deep ensemble + calibration | 85.8% | 14.2% | 89.6% | 0.044 | 4.4% |
| Ensemble + conformal selective deferral | 77.6% | 22.4% | 94.1% | 0.028 | 2.1% |

**Source:** Author's synthetic simulation. Values do not represent an existing administrative system.

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

### 5.2 Safety–Coverage Trade-Off

The analysis indicates that reliability was not achieved by increasing predictive accuracy alone. It was achieved partly by limiting the circumstances in which the system was permitted to act automatically.

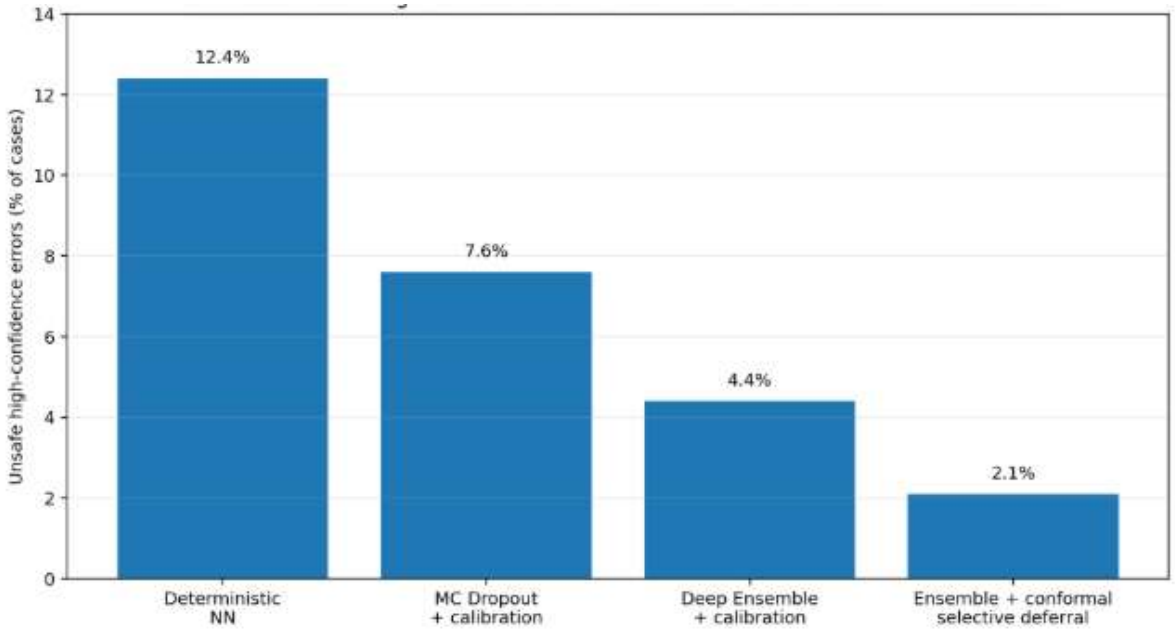
The deterministic architecture produced maximum administrative coverage but also the highest rate of high-confidence errors. Introducing MC dropout reduced automatic processing modestly but intercepted a portion of uncertain cases. Deep ensembles reduced the unsafe-error rate further by incorporating disagreement among independently trained models.

The conformal selective architecture produced the strongest simulated safety outcome but required the largest review workload. Approximately one in every five cases was redirected away from automatic processing. This result demonstrates why uncertainty-aware administrative design should be understood as a resource-allocation problem as well as a machine-learning problem.

A very conservative uncertainty threshold can reduce automated error but may overwhelm human reviewers and undermine efficiency. A permissive threshold preserves throughput but may allow more unjustifiably confident mistakes. The appropriate threshold therefore depends on the consequences of error, available review resources, reversibility of decisions, and the legal or procedural rights affected.

### 5.3 Unsafe High-Confidence Errors

**Graph 1: Simulated Unsafe High-Confidence Error Rate Across Neural Decision Architectures**




The graph demonstrates a progressive decline in incorrect decisions that remain both highly confident and unflagged by the decision architecture. The baseline rate of 12.4% falls to 7.6% with Monte Carlo dropout and calibration, 4.4% under the calibrated deep ensemble, and 2.1% under the ensemble-plus-conformal selective framework.The improvement should not be interpreted as proof that conformal or ensemble systems would produce the same reduction in a real agency. The values were deliberately generated to demonstrate an evaluation architecture.

The more important methodological finding is that conventional accuracy should not be the sole metric for high-stakes administrative AI. A system may exhibit respectable overall accuracy while still generating a problematic subgroup of confident, difficult-to-detect mistakes. Measuring high-confidence error, calibration, coverage, and deferral provides a more informative picture of the model's suitability for consequential decision support.

## 6. Discussion

### 6.1 Uncertainty Should Determine the Level of Administrative Automation

The simulation demonstrates why uncertainty should be treated as an operational boundary on automation rather than as a technical statistic displayed after a decision has already been produced. A deterministic neural classifier implicitly behaves as though every case deserves a definitive answer. This assumption is poorly suited to administrative environments in which cases vary substantially in documentary quality, factual complexity, similarity to previous cases, and potential consequences of error. Uncertainty-aware architectures permit a more proportionate relationship between algorithmic confidence and governmental action. Routine cases that closely resemble validated examples can receive greater computational assistance, whereas unfamiliar or ambiguous cases can be escalated before an automated output acquires legal or practical effect.

This principle corresponds with NIST's risk-management approach, which recommends that organizations identify risk tolerance, allocate stronger controls where tolerances are lower, and continuously revisit risk-management decisions as new information becomes available. In administrative applications, the cost of uncertainty should therefore be interpreted in relation to the consequence of error. A minor prioritization recommendation may justify a different threshold from an automated recommendation capable of suspending an essential benefit or substantially influencing immigration status. The threshold for automatic processing should become more conservative as the potential harm becomes more severe, difficult to reverse, or difficult for the affected person to identify.

The same logic appears in public-sector governance outside machine-learning research. Canada's Algorithmic Impact Assessment evaluates automated systems according to features including the nature of the decision and its potential impact, while the Council of Europe's HUDERIA approach considers variables such as scale, scope, reversibility, and likelihood. An uncertainty-governed model extends these governance principles into the decision architecture itself by making automatic resolution conditional on the model satisfying a contextually determined reliability threshold.

### 6.2 Human Review Must Be Designed for Uncertain Cases Rather Than Added Symbolically

Selective deferral is useful only when the receiving human process is capable of correcting what the model could not resolve. Merely transferring low-confidence cases to an official who sees the same algorithmic recommendation and routinely approves it would provide little substantive protection. Human review should therefore be designed specifically around the source of uncertainty. If uncertainty results from incomplete documentation, the reviewer should be able to request additional evidence. If the model detects an unfamiliar pattern, the reviewer should have access to contextual information not represented in the model. If multiple outcomes remain statistically plausible, the reviewer should be

informed that the case was deferred because the model lacked sufficient certainty rather than being shown an apparently authoritative recommendation without qualification.

This distinction is especially important because Canada explicitly treats automated systems as subject to its directive even when a human officer formally makes the final decision. NIST similarly recognizes that human–AI arrangements span a continuum from autonomous decision-making to human use of AI as an additional opinion. The relevant governance question is therefore not whether a person appears somewhere in the workflow but whether the architecture enables that person to exercise meaningful independent judgment.

Human review also introduces its own uncertainty and potential inconsistency. The purpose of selective automation should not be to idealize human decision-making as error-free. Instead, the goal is to allocate cases according to comparative strengths. Neural networks may efficiently process regular patterns across large datasets, whereas trained officials may be better positioned to interpret unusual circumstances, evaluate explanations, reconcile conflicting evidence, and apply discretion. A mature system should therefore evaluate not only model accuracy but also the outcomes of deferred cases, disagreement between humans and models, reversal rates, time to resolution, and whether reviewers systematically defer to algorithmic outputs.

### 6.3 Calibration, Distribution Shift, Fairness, and Public Accountability

Good average calibration does not guarantee good calibration for every population group. A model may be well calibrated overall while becoming overconfident for a smaller demographic, geographic, or socioeconomic subgroup. This concern is particularly important in public administration because unequal uncertainty can translate into unequal exposure to automated error or unequal rates of referral to slower human procedures. The governance framework should therefore evaluate calibration, uncertainty, coverage, and false-confidence rates at appropriate subgroup levels where lawful, statistically meaningful, and ethically justified.

Distribution shift creates a related problem. Ovadia and colleagues demonstrated that uncertainty-estimation methods can deteriorate as deployment data diverge from the development distribution, even when sophisticated methods are used. Consequently, an uncertainty-aware model cannot be certified once and then assumed to remain reliable indefinitely. Post-deployment monitoring should examine input drift, calibration drift, changes in uncertainty distributions, changes in deferral volumes, emerging error clusters, and model performance following policy or population changes. NIST's AI RMF likewise treats AI risk management as a lifecycle process rather than a one-time technical assessment.

Transparency should also include uncertainty governance itself. The UK's Algorithmic Transparency Recording Standard is designed to provide meaningful public information concerning how and why algorithmic systems are used in public decision-making. For a high-stakes uncertainty-aware system, meaningful transparency should therefore describe not only the existence of a neural model but the role it plays, whether it can produce automatic outcomes, the circumstances that lead to deferral, how model performance is monitored, and what happens when uncertainty is high. It would be misleading for a public authority to describe a system simply as “95% accurate” without explaining that performance

may vary by case type, that uncertainty thresholds limit its autonomous use, or that some applicant categories are systematically referred for manual review.

The current EU AI Act reinforces the importance of this broader governance approach through requirements for risk management, human oversight, and appropriate levels of accuracy and robustness in high-risk systems. For the Annex III high-risk categories covered by the revised timetable, those high-risk rules are scheduled to apply from December 2, 2027 following the July 2026 AI Omnibus. Uncertainty-aware model design should therefore be viewed not as a substitute for regulatory compliance but as one technical mechanism that can support a wider system of risk management, monitoring, documentation, human oversight, and accountability.

## 7. Conclusion

Artificial intelligence can improve the capacity of public agencies to process large and complex administrative workloads, but high predictive accuracy alone is an inadequate basis for high-stakes administrative automation. A neural network may be accurate on average and still produce seriously misleading high-confidence predictions, particularly when it encounters ambiguous records or conditions that differ from those represented during model development. The central contribution of uncertainty-aware machine learning is therefore not simply that it attaches an additional numerical score to a prediction. Its deeper contribution is the ability to recognize circumstances in which the system should refrain from acting with unwarranted certainty.

The simulation presented in this study illustrates this principle through four progressively uncertainty-aware architectures. The deterministic neural network achieved complete automatic coverage but produced a simulated unsafe high-confidence error rate of 12.4%. Monte Carlo dropout and calibration reduced that rate to 7.6%, while the calibrated deep ensemble reduced it to 4.4%. The ensemble-plus-conformal selective architecture produced the lowest simulated rate at 2.1%, but automatically resolved only 77.6% of cases because 22.4% were referred for additional review. These figures are synthetic and must not be cited as real-world effectiveness estimates. Their purpose is to demonstrate the safety–coverage trade-off that should be measured when uncertainty-aware models are evaluated for consequential administrative use.

The results also demonstrate that uncertainty-aware AI is fundamentally a governance problem. An agency must determine how much uncertainty is tolerable for a particular decision, what consequences follow from a deferral, who reviews uncertain cases, how quickly those cases are resolved, whether uncertainty differs across population groups, how changing data distributions are detected, and how citizens can question or correct an outcome. A technically sophisticated uncertainty estimator can still produce an unjust administrative process if low-confidence cases are routed into an under-resourced review system or if uncertainty disproportionately burdens particular groups.

A defensible model for high-stakes administration is therefore uncertainty-governed automation rather than unrestricted prediction. Under this model, neural systems may automate or strongly support cases only when their calibrated uncertainty falls within a predefined, context-sensitive risk tolerance. Cases exceeding that threshold should trigger additional evidence gathering, independent human assessment, or another structured administrative safeguard. Model calibration, deferral rates, high-confidence errors, distribution shift, subgroup reliability, and outcomes of human review should then be

monitored throughout the system lifecycle. Such an approach does not eliminate error or convert uncertainty into certainty. It does, however, replace the dangerous assumption that an algorithm must always know the answer with a more appropriate principle for public administration: when the system does not know enough, that uncertainty should change what government is permitted to do next.

**References**

1. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 1050–1059.
2. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems, 30*, 5574–5584.
3. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems, 30*, 6402–6413.
4. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*, 1321–1330.
5. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems, 32*, 13991–14002.
6. Malinin, A., & Gales, M. (2018). Predictive uncertainty estimation via prior networks. *Advances in Neural Information Processing Systems, 31*, 7047–7058.
7. Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural networks. *Proceedings of the 32nd International Conference on Machine Learning*, 1613–1622.
8. Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer.
9. MacKay, D. J. C. (1992). A practical Bayesian framework for backpropagation networks. *Neural Computation, 4*(3), 448–472. https://doi.org/10.1162/neco.1992.4.3.448
10. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
11. Kendall, A., Badrinarayanan, V., & Cipolla, R. (2015). Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*.
12. Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems, 31*, 3179–3189.
13. Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations (ICLR)*.
14. Liang, S., Li, Y., & Srikant, R. (2018). Enhancing the reliability of out-of-distribution image detection in neural networks. *International Conference on Learning Representations (ICLR)*.
15. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
16. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

17. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774.
18. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
19. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. https://doi.org/10.1145/3287560.3287598
20. Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, *165*(3), 633–705.