Advanced AI OHS Professional Certification (Part II) (Canada)

Applying AI in Incident Investigation and Root Cause Analysis

Anomaly Detection (Related: Outlier Detection, Pattern Recognition) – A computational technique that identifies data points or events that deviate markedly from established norms. In incident investigation, anomaly detection algorithms scan sensor streams, log files, or safety reports to flag unusual spikes that may precede a hazard. For example, a sudden increase in temperature readings on a production line can be automatically highlighted for review. Challenges include setting appropriate sensitivity thresholds to avoid excessive false alarms and ensuring the algorithm adapts to evolving operational baselines without manual recalibration.

Artificial Intelligence (AI) (Related: Machine Learning, Cognitive Computing) – The broader discipline encompassing systems that mimic human intelligence to perform tasks such as reasoning, learning, and problem solving. In the context of occupational health and safety, AI can ingest large volumes of incident data, learn patterns, and suggest preventive measures. A key challenge is maintaining data quality and representativeness; biased or incomplete records can lead to skewed insights and undermine trust among safety professionals.

Bias (Related: Data Bias, Algorithmic Fairness) – Systematic error introduced into AI outputs due to skewed training data, flawed model assumptions, or unbalanced feature representation. In root cause analysis, bias may cause the system to over-emphasize certain incident types while neglecting others, potentially misdirecting corrective actions. Mitigation strategies include diverse data collection, bias audits, and incorporating fairness constraints during model development.

Causal Inference (Related: Correlation, Counterfactual Analysis) – A statistical framework that moves beyond mere association to identify cause-and-effect relationships. When AI suggests that “increased overtime correlates with higher injury rates,” causal inference techniques test whether overtime itself contributes to risk or merely co-occurs with other factors. Implementing robust causal models often requires longitudinal data and careful control of confounding variables, which can be resource-intensive.

Deep Learning (Related: Neural Networks, Representation Learning) – A subset of machine learning that employs multi-layered artificial neural networks to automatically extract high-level features from raw data. Convolutional neural networks (CNNs) can analyze video footage of workplace incidents to detect unsafe postures, while recurrent neural networks (RNNs) process time-series sensor data for early warning signals. The primary challenge is the need for large labeled datasets and significant computational power, which may be prohibitive for smaller organizations.

Explainable AI (XAI) (Related: Model Transparency, Interpretability) – Techniques that make AI decision processes understandable to human users. In incident investigations, XAI can provide a visual trace of why a model flagged a particular machine as high-risk, highlighting contributing variables such as vibration

intensity and maintenance history. Balancing explanatory depth with model performance is a common trade-off; overly simplified explanations may miss nuance, while complex ones can overwhelm end-users.

Feedback Loop (Related: Continuous Improvement, Model Retraining) – The cyclical process where outcomes of AI-driven recommendations are monitored, and the resulting data are fed back into the model to refine predictions. For instance, after implementing a suggested engineering control, the reduction in incident frequency is logged and used to adjust the model's risk weighting. Managing feedback loops requires disciplined data governance to prevent drift caused by erroneous or incomplete post-implementation data.

Ground Truth (Related: Labeled Data, Validation Set) – The definitive, verified information against which AI predictions are measured. In safety analytics, ground truth may consist of manually reviewed incident reports confirmed by safety officers. High-quality ground truth is essential for training reliable models; however, acquiring it can be labor-intensive, especially when historical records are incomplete or inconsistently formatted.

Human-In-The-Loop (HITL) (Related: Decision Support, Oversight) – A design paradigm where AI assists but does not replace human judgment. During root cause analysis, the system proposes probable causes, and safety professionals validate, adjust, or reject these suggestions. HITL safeguards against over-reliance on opaque algorithms but introduces potential bottlenecks if human review becomes a delay point. Effective HITL implementation balances automation speed with necessary expert oversight.

Incident Investigation (Related: Accident Analysis, Event Reconstruction) – A systematic process of examining the sequence of events leading to a workplace incident to identify immediate and underlying causes. AI augments this process by rapidly aggregating evidence, correlating similar past incidents, and surfacing hidden patterns. A challenge lies in integrating AI outputs with traditional narrative reports while preserving the investigative rigor required by regulatory bodies.

Knowledge Graph (Related: Semantic Network, Ontology) – A structured representation of entities (e.G., Equipment, procedures, personnel) and their interrelationships. In safety contexts, a knowledge graph can link a malfunctioning valve to its maintenance schedule, associated training modules, and prior incident records. Populating and maintaining an accurate graph demands ongoing curation and alignment with evolving organizational terminology.

Machine Learning (ML) (Related: Supervised Learning, Feature Engineering) – A collection of algorithms that enable computers to learn patterns from data without explicit programming. In OHS, ML models predict injury likelihood based on variables such as shift length, equipment age, and environmental conditions. Key challenges include feature selection bias, overfitting to historical trends that may not hold under new operational regimes, and ensuring model updates reflect current practices.

Natural Language Processing (NLP) (Related: Text Mining, Sentiment Analysis) – Techniques for extracting meaning from unstructured textual data. NLP can automatically parse incident narratives, extract key factors (e.G., "Slipped on wet floor"), and classify severity levels. Ambiguities in language, domain-specific jargon, and multilingual reports pose significant hurdles, often requiring custom tokenization and domain-adapted

language models.

OpenAI (Related: Large Language Models, API Integration) – An organization that develops advanced generative AI models, such as GPT series, which can be leveraged to draft investigation summaries, suggest corrective actions, or answer safety-related queries. While these models excel at language generation, they may hallucinate facts or lack domain-specific accuracy, necessitating rigorous verification before adoption in formal reports.

Predictive Analytics (Related: Forecasting, Risk Scoring) – The practice of using statistical techniques and AI models to anticipate future events based on historical data. In OHS, predictive analytics can assign a risk score to each worksite, prompting proactive inspections. Limitations arise from data latency, changing regulatory landscapes, and the risk of over-reliance on numerical scores without contextual interpretation.

Quantitative Risk Assessment (QRA) (Related: Probabilistic Modeling, Hazard Quantification) – A systematic method that assigns numerical probabilities and consequences to identified hazards. AI can automate the aggregation of incident frequencies, exposure durations, and severity factors to compute QRA metrics. However, quantitative models may oversimplify complex sociotechnical interactions, and data scarcity for rare events can lead to unreliable probability estimates.

Root Cause Analysis (RCA) (Related: Five Whys, Fault Tree Analysis) – A disciplined approach to uncovering the fundamental origins of an incident, rather than merely addressing symptoms. AI-enhanced RCA leverages pattern mining to suggest plausible root causes based on similarity to past cases. The principal challenge is ensuring that AI suggestions do not bypass critical human reasoning, especially when causal chains involve intangible factors such as safety culture.

Supervised Learning (Related: Labeled Training Data, Classification) – A machine-learning paradigm where models are trained on input-output pairs, learning to map features to known outcomes. For injury prediction, labeled examples of “safe” versus “unsafe” conditions enable the model to classify new observations. Obtaining comprehensive labeled datasets is often a bottleneck, and class imbalance (few incidents vs. Many safe observations) can degrade model performance.

Temporal Data Mining (Related: Time-Series Analysis, Sequence Mining) – Techniques for extracting patterns from chronologically ordered data. In incident investigation, temporal mining can reveal that equipment failures tend to cluster after specific maintenance intervals, informing schedule adjustments. Handling irregular sampling rates, missing timestamps, and seasonal variations requires sophisticated preprocessing and model selection.

Unsupervised Learning (Related: Clustering, Anomaly Detection) – Algorithms that identify inherent structure in data without predefined labels. Clustering incident records can uncover hidden groups, such as “near-misses involving manual handling” that share common attributes. The interpretability of clusters can be ambiguous, and selecting the appropriate number of clusters often relies on domain expertise.

Validation Set (Related: Test Set, Cross-Validation) – A subset of data reserved for tuning model hyperparameters and assessing generalization performance before final testing. Proper validation prevents over-optimistic estimates of AI efficacy in safety contexts. Care must be taken to ensure temporal

separation (e.G., Using older incidents for training and newer ones for validation) to mimic real-world forecasting conditions.

Weighted Scoring Model (Related: Multi-Criteria Decision Analysis, Prioritization) – A framework that assigns weights to various risk factors (e.G., Exposure frequency, severity potential) to compute an overall hazard score. AI can dynamically adjust weights based on observed incident outcomes, delivering a data-driven prioritization scheme. Determining appropriate weights remains subjective, and frequent recalibration may cause score volatility, confusing stakeholders.

Explainable Risk Dashboard (Related: Visualization, Decision Support) – An interactive interface that presents AI-derived risk indicators alongside explanatory notes, allowing safety officers to explore why a particular hazard received a high rating. Effective dashboards combine clear visual cues with concise textual explanations. Over-crowding the interface or presenting overly technical jargon can impede user adoption.

Zero-Shot Learning (Related: Transfer Learning, Few-Shot Learning) – An advanced AI capability where a model can recognize or classify instances of a category it has never seen during training, based on semantic descriptions. In OHS, a zero-shot model might identify a novel type of equipment malfunction by relating it to known failure modes. The primary limitation is reliance on high-quality semantic embeddings; inaccurate descriptions can lead to misclassification.

Algorithmic Transparency (Related: Model Documentation, Audit Trail) – The practice of openly documenting AI model architecture, training data sources, and decision logic. Transparency builds trust among regulators and workers when AI suggests corrective actions. However, fully disclosing proprietary model details may conflict with commercial confidentiality, requiring a balance between openness and intellectual property protection.

Bayesian Network (Related: Probabilistic Graphical Model, Causal Modeling) – A directed acyclic graph where nodes represent variables and edges encode probabilistic dependencies. In incident analysis, a Bayesian network can model the likelihood that “lack of PPE” leads to “injury severity,” updating beliefs as new evidence arrives. Constructing accurate conditional probability tables demands expert elicitation, and the networks can become computationally intensive with many variables.

Change Management (Related: Organizational Readiness, Adoption Strategy) – The structured approach to transitioning people, processes, and technology when introducing AI tools into safety programs. Successful change management includes stakeholder communication, training on AI interpretation, and clear governance policies. Resistance may arise from fear of job displacement or skepticism about algorithmic accuracy, necessitating transparent benefit demonstrations.

Data Governance (Related: Data Quality, Privacy Compliance) – The set of policies, standards, and procedures that ensure data used for AI is accurate, secure, and ethically managed. In OHS, governance must address confidentiality of personal injury records, retention schedules, and consent for data sharing. Weak governance can result in regulatory penalties, loss of employee trust, and compromised model reliability.

Edge Computing (Related: IoT, Real-Time Analytics) – Processing data near the source (e.G., On-site sensors)

rather than transmitting it to centralized servers. Edge AI can deliver instant hazard detection, such as identifying a worker entering a restricted zone, without latency. Constraints include limited computational resources on edge devices and challenges in updating models across distributed hardware.

Federated Learning (Related: Distributed Training, Privacy-Preserving AI) – A technique where multiple organizations train a shared model locally on their own data, sending only model updates to a central server. This approach enables collaborative safety analytics across companies while keeping proprietary incident data private. The trade-off includes increased communication overhead and potential heterogeneity in data quality across participants.

Generative Adversarial Network (GAN) (Related: Synthetic Data Generation, Model Augmentation) – A pair of neural networks that compete to produce realistic synthetic data. In safety research, GANs can generate plausible incident scenarios to augment scarce training sets for rare event prediction. Synthetic data may inadvertently embed biases present in the original dataset, and evaluating realism requires domain expert validation.

Human Factors Engineering (Related: Ergonomics, System Design) – The discipline that studies how people interact with equipment, environments, and procedures, aiming to reduce error and improve safety. AI can incorporate human-factor variables (e.G., Fatigue scores) into risk models, but quantifying such subjective factors remains difficult, often relying on indirect proxies like shift length.

Incident Severity Index (Related: Risk Matrix, Impact Scoring) – A numerical composite that reflects both the frequency and consequence of reported incidents. AI can automatically compute the index by extracting injury type, lost-time days, and medical costs from reports. Calibration of the index is crucial; over-weighting financial cost may downplay psychological injuries, leading to incomplete safety priorities.

Just-In-Time (JIT) Alerts (Related: Real-Time Notification, Proactive Intervention) – Immediate warnings delivered to workers or supervisors when AI detects a developing unsafe condition, such as abnormal vibration levels on machinery. The effectiveness of JIT alerts hinges on minimizing false positives, as excessive alerts can cause “alarm fatigue” where users start ignoring messages.

K-Means Clustering (Related: Unsupervised Learning, Partitioning Algorithm) – A simple algorithm that groups data points into k clusters based on distance to cluster centroids. Applied to incident logs, K-means can reveal clusters of similar root causes, aiding targeted training. Choosing the correct k value often requires iterative testing and domain insight; inappropriate k leads to over- or under-segmentation.

Latent Variable Model (Related: Hidden Factors, Dimensionality Reduction) – Models that infer unobserved variables that explain observed data patterns. Factor analysis can uncover latent safety culture dimensions influencing incident rates. Estimating latent variables demands robust statistical techniques and sufficient data; sparse or noisy datasets can produce unstable factor loadings.

Model Drift (Related: Concept Drift, Performance Degradation) – The gradual decline in AI model accuracy as underlying data distributions change over time, such as new equipment introductions altering incident patterns. Detecting drift involves monitoring prediction error metrics and triggering retraining. Ignoring drift can lead to misleading risk assessments and erode stakeholder confidence.

Neural Architecture Search (NAS) (Related: AutoML, Model Optimization) – Automated methods that explore different neural network configurations to identify the most effective architecture for a given task. In safety analytics, NAS can discover lightweight models suitable for edge devices. The process is computationally expensive and may produce architectures that are difficult to interpret, conflicting with explainability requirements.

Ontology Alignment (Related: Semantic Integration, Knowledge Graph Merging) – The process of reconciling differing vocabularies and concepts across multiple data sources. Aligning a contractor's incident taxonomy with a host organization's ontology enables unified AI analysis. Misalignment can cause duplicate entities or loss of nuance, reducing the fidelity of cross-organizational insights.

Pattern Mining (Related: Association Rules, Frequent Itemsets) – Techniques that discover recurring combinations of variables within large datasets. In incident records, pattern mining may reveal that "use of ladder + wet floor" frequently precedes falls. The volume of generated rules can be overwhelming; pruning based on statistical significance and actionable relevance is essential.

Quantile Regression (Related: Predictive Modeling, Distributional Forecasting) – A regression method that estimates conditional quantiles (e.G., Median, 95th percentile) of a response variable, providing a fuller picture of risk variability. Applying quantile regression to injury cost data can help planners understand worst-case financial exposure. The method assumes sufficient data across the distribution; sparse tail data can produce unstable quantile estimates.

Reinforcement Learning (Related: Policy Optimization, Agent-Environment Interaction) – An AI paradigm where an agent learns optimal actions through trial-and-error feedback from its environment. In safety simulations, a reinforcement-learning agent can explore different preventive strategies and learn which actions minimize incident occurrence. Real-world deployment is limited by the need for safe exploration; unsafe trial actions are not permissible in live workplaces.

Scenario Analysis (Related: What-If Modeling, Stress Testing) – The systematic evaluation of potential future events by varying key inputs to assess impacts on safety outcomes. AI can generate thousands of plausible incident scenarios based on stochastic models, helping organizations test the robustness of control measures. The quality of scenarios depends on the realism of underlying probability distributions; unrealistic assumptions can mislead decision-makers.

Time-Series Forecasting (Related: ARIMA, LSTM) – Predictive techniques that project future values of a metric based on its historical temporal pattern. Forecasting monthly lost-time injury rates helps allocate inspection resources. Complex seasonality, external shocks, and non-stationarity present modeling challenges; advanced deep-learning models may mitigate but also increase interpretability concerns.

Uncertainty Quantification (Related: Confidence Intervals, Probabilistic Outputs) – The practice of explicitly measuring the degree of confidence in model predictions. Providing safety managers with a probability range (e.G., 70-90 % Confidence that a hazard exceeds threshold) supports risk-aware decision making. Quantifying uncertainty requires appropriate statistical techniques and may increase computational load.

Variable Importance (Related: Feature Attribution, Sensitivity Analysis) – Metrics that indicate how much

each input contributes to a model's prediction. In injury risk models, variable importance can reveal that "night shift" carries more weight than "age." Importance scores can be misleading if features are correlated; advanced methods like SHAP values help disentangle joint effects.

Weighted Ensemble (Related: Model Stacking, Hybrid Modeling) – Combining predictions from multiple models, each assigned a weight based on performance, to improve overall accuracy. An ensemble of decision trees, logistic regression, and neural networks can capture diverse patterns in incident data. Determining optimal weights requires validation, and ensembles can become opaque, challenging explainability mandates.

Zero-Trust Architecture (Related: Cybersecurity, Access Control) – A security framework that requires continuous verification of user identity and device integrity before granting access to AI systems and data. Implementing zero-trust safeguards sensitive incident records from unauthorized exposure. The architecture adds complexity to system integration and may impact latency for real-time analytics.