Canadian Journal of Machine Learning & Intelligent Transportation (2026), Vol. **1**
https://doi.org/10.65269/3twm3e15

ARTICLE

From Predictive Accuracy to Human-Centric Decision Support: An Operational HCT-ML Framework for Intelligent Transportation Systems

Bappa Muktar *

Department of Computer Science, Université du Québec en Outaouais (UQO), 283 Boul. Alexandre-Taché, Gatineau, QC J8X 3X7, Canada

*Corresponding author: mukb06@uqo.ca

(Received: 22 July 2026; Revised: 15 August 2026; Accepted: 04 September 2026; Published: 05 September 2026)
(Editor: Dr. Tarek Othmani; Reviewers: Anonymous reviewers)

Abstract

Machine learning in intelligent transportation systems (ITS) is commonly evaluated through predictive performance, yet deployment decisions also depend on whether outputs are relevant to a defined decision, understandable to intended users, equitable across affected groups, uncertainty-aware, subject to appropriate human authority, and actionable within operational constraints. This Perspective develops the Human-Centric and Trustworthy Machine Learning (HCT-ML) framework as an operational decision-support profile for ITS. Its novelty is not the invention of new responsible artificial intelligence (AI) principles; rather, HCT-ML integrates established requirements around four transport-specific constructs—decision owner, decision horizon, intervention pathway, and consequence of error—and translates six human-centric dimensions (relevance, explainability, fairness, uncertainty, human oversight, and actionability) into evidence requirements, candidate indicators, context-specific thresholds, and non-compensatory deployment gates. The framework is benchmarked against the National Institute of Standards and Technology (NIST) AI Risk Management Framework, Organisation for Economic Co-operation and Development (OECD) AI Principles, the European Union (EU) AI Act, Institute of Electrical and Electronics Engineers (IEEE) 7000-series standards, and United States Department of Transportation (USDOT) AI-assurance guidance. We further provide a formal operationalization, application-specific priority profiles, real-world safety cases from automated-driving investigations, and a worked municipal road-safety protocol using the Montréal open-data context. The article does not claim empirical validation or causal safety gains; instead, it provides an evidence-informed and reproducible protocol that can be tested through stakeholder studies, simulations, and field deployments.

Keywords: Human-Centered Artificial Intelligence; Intelligent Transportation Systems; Machine Learning; Road Safety; Trustworthy Artificial Intelligence

Abbreviations: ADS: Automated driving system; AI: Artificial intelligence; AI RMF: Artificial Intelligence Risk Management Framework; EU: European Union; HCT: Human-Centric and Trustworthy; HCT-ML: Human-Centric and Trustworthy Machine Learning; IEEE: Institute of Electrical and Electronics Engineers; ITS: Intelligent transportation systems; NIST: National Institute of Standards and Technology; NTSB: National Transportation Safety Board; OECD: Organisation for Economic Co-operation and Development; OOD: Out-of-distribution; SHIFT: A Synthetic Driving Dataset for Continuous Multi-Task Domain Adaptation; USDOT: United States Department of Transportation

1. Introduction

Machine learning has become a central component of intelligent transportation systems (ITS), supported by traffic sensors, connected vehicles, mobile devices, and increasingly accessible urban data. ITS research encompasses traffic management, autonomous vehicles, mobility prediction, and other smart-city applications [1]. Predictive models are also used for traffic-collision prediction [2], travel and arrival-time forecasting [3], driver-state and fatigue monitoring [4], and driving-behaviour analysis [5]. Progress across these applications is commonly expressed through gains in accuracy, precision, recall, or reductions in prediction error. These measures are indispensable for technical comparison, but their prominence has encouraged a narrow account of what makes a transportation model successful.

Predictive performance alone does not establish that a system is useful, safe, or socially acceptable. A highly accurate model may be difficult for traffic managers to interpret, perform unevenly across road-user groups or neighbourhoods, or provide overconfident outputs under unfamiliar conditions. Its practical value may also remain limited when a prediction does not clarify what action should follow or how uncertainty should influence that action. These weaknesses are consequential in road-safety settings, where model outputs can affect infrastructure priorities, operational interventions, public warnings, and emergency responses. Human-centered artificial intelligence (AI) consequently calls for systems that combine computational capability with meaningful human control, transparency, and sensitivity to context [6]. Research on explainable artificial intelligence similarly stresses that explanations should be evaluated according to the needs of intended users rather than treated only as technical properties of a model [7].

This Perspective argues that machine learning for ITS should be evaluated not only by how accurately it predicts an event, but also by how effectively it supports a defined human or institutional decision. Human-Centric and Trustworthy Machine Learning (HCT-ML) is deliberately positioned as a transport-specific operational profile that complements, rather than replaces, established responsible-AI and assurance frameworks. Its added value lies in making the decision interface explicit: who owns the decision, when the decision must be made, what intervention is available, which groups bear the consequences, and what should happen when confidence is insufficient. In this sense, HCT-ML narrows broad principles into a protocol for evaluating a transportation prediction at the point where it becomes an operational action.

The article makes four contributions. First, it benchmarks HCT-ML against the National Institute of Standards and Technology (NIST), the Organisation for Economic Co-operation and Development (OECD), the European Union (EU), the Institute of Electrical and Electronics Engineers (IEEE), and the United States Department of Transportation (USDOT) guidance and states the boundaries of the six selected dimensions. Second, it introduces a formal operationalization in which each dimension is supported by measurable evidence and minimum context-specific gates rather than by an unconstrained weighted score. Third, it provides transport-application priority profiles and explicit rules for documenting trade-offs. Fourth, it adds a case-based plausibility check using real automated-driving safety investigations and a worked Montréal municipal road-safety protocol. These additions do not constitute empirical validation; they specify what future validation must measure and prevent the conceptual framework from being presented as a demonstrated safety intervention.

The remainder of the article explains why accuracy-centered evaluation is insufficient, examines the principal human stakeholders, derives and operationalizes HCT-ML, demonstrates how the framework can be applied to transportation decisions, and outlines an empirical validation and deployment agenda.

2. Why Accuracy-Centric Machine Learning Is Insufficient

Accuracy-centered evaluation is useful for comparing models, but it provides only a partial account of their suitability for intelligent transportation systems. Road-safety datasets can contain substantial class imbalance, allowing strong aggregate accuracy to coexist with weak recognition of underrepresented outcomes [2]. In congestion and public transport forecasting, low average error can likewise conceal failures during disruptions, extreme weather, or unusual demand.

The first limitation is interpretability. Many high-performing models provide little insight into why a collision, traffic condition, or driving manoeuvre has been classified as hazardous. In autonomous driving, a correct control action is insufficient when engineers, safety assessors, or users cannot identify the environmental cues that shaped it. Explainability is therefore relevant to regulatory compliance and social acceptance in autonomous driving [8]. A municipal risk score is similarly difficult to use when professionals cannot determine whether it reflects road geometry, weather, exposure, behaviour, or an artefact of the data.

A second limitation is the gap between offline evaluation and real-world use. Historical test sets represent bounded conditions, whereas deployed systems encounter sensor failures, construction zones, changing mobility patterns, unfamiliar road layouts and weather conditions that were underrepresented in the training data. The benchmark *A Synthetic Driving Dataset for Continuous Multi-Task Domain Adaptation* (SHIFT) illustrates how changes in illumination, precipitation, traffic density, and pedestrian activity can degrade autonomous-driving perception [9]. Robustness, calibration, latency, and distribution shift must therefore be assessed throughout deployment.

Aggregate accuracy may also conceal unequal outcomes. Uneven sensing, incomplete reporting, and historically unequal access to transport services can cause some communities to be represented more reliably than others. Travel-behaviour research has documented disparities associated with ethnicity, income, disability, and geography [10]. Fairness-aware travel-demand forecasting likewise shows that accurate aggregate predictions can coexist with biases across protected attributes [11]. When predictions guide service allocation, infrastructure investment, or enforcement, modelling disparities can become policy disparities.

Finally, models are often developed without sufficient attention to those responsible for acting on them. Traffic controllers, transit operators, emergency personnel, engineers, and planners work within legal and operational constraints that cannot be reduced to an objective function. They must know when to trust or override a recommendation and who remains accountable. Responsible AI frameworks therefore treat governance and human oversight as lifecycle requirements [12]. The relevant question is not only whether a model predicts accurately, but whether it enables informed, equitable, and accountable action.

2.1 Real-world evidence that technical performance is not sufficient

Transportation accident investigations provide concrete evidence that system safety cannot be inferred from an algorithmic or automation-performance metric in isolation. In the 2018 Tempe, Arizona, fatal crash involving a developmental automated driving system (ADS), the United States National Transportation Safety Board (NTSB) identified not only the vehicle operator's failure to monitor the environment, but also inadequate safety-risk assessment, ineffective oversight of operators, and insufficient mechanisms for automation complacency as contributing organizational factors [13]. In the 2018 Mountain View, California, crash involving partial driving automation, the NTSB identified system limitations, driver overreliance, and ineffective monitoring of driver engagement among the safety issues [14]. These investigations do not establish that a highly accurate machine-learning predictor caused or could have prevented either crash. They do, however, show why technical capability alone is an incomplete basis for safety claims: limitations must be commu-

nicated, fallback authority must be effective, human attention must be realistically supported, and operational governance must anticipate foreseeable misuse and complacency.

Table 1 summarizes how these investigations map to HCT-ML dimensions and highlights the system-level requirements that predictive performance alone does not capture.

Table 1. Real-world transportation safety cases illustrating system-level requirements beyond predictive performance.

Case	Investigation finding relevant to decision support	HCT-ML dimensions highlighted	Why prediction/performance alone is insufficient
Tempe developmental ADS, 2018 [13]	NTSB identified inadequate safety-risk assessment, ineffective operator oversight, and automation complacency among contributing factors.	Human oversight; relevance; actionability; uncertainty	A system requires credible fallback roles, monitoring, and risk controls even when automated perception and control are technically capable.
Mountain View partial automation, 2018 [14]	NTSB identified automation-system limitations, driver overreliance, and ineffective driver-engagement monitoring.	Explainability; uncertainty; human oversight; actionability	Users must understand system limits and be able to intervene within the available time window; nominal automation performance does not guarantee safe use.

These cases motivate HCT-ML as a decision-support evaluation layer, not as a claim that the six dimensions are themselves proven causal determinants of safety. The safety contribution proposed in this article is therefore a documented mechanism: each dimension is tied to a failure mode, an observable indicator, and a corresponding mitigation or fallback action.

Taken together, these limitations suggest that stakeholder-specific requirements, rather than aggregate model performance alone, should provide the starting point for the design and evaluation of transportation decision-support systems.

3. Human Stakeholders in Intelligent Transportation

Intelligent transportation systems operate within a network of people and institutions whose needs cannot be represented by one performance objective. Some stakeholders interact directly with predictive technologies, while others experience the consequences of decisions based on their outputs. Human-centered design must therefore consider both who uses the model and who bears the risk when it is incomplete or wrong.

Drivers receive collision warnings, route recommendations, fatigue alerts, and automated driving assistance. Information must arrive in time to support action without increasing distraction or cognitive load. A technically accurate warning may still fail if its meaning is unclear, its timing is poor, or repeated false alarms encourage disregard. Pedestrians and cyclists often do not interact with the model directly, yet vehicle perception, signal control, and infrastructure prioritization can affect them substantially. This concern is especially important because vulnerable road users account for more than half of global road deaths [15].

People with reduced mobility have additional requirements concerning accessibility, reliability, and independence. Route planning, demand-responsive services, automated vehicles, and passenger information must account for physical, sensory, and cognitive constraints obscured by average-user assumptions. Transport policy should therefore examine access to opportunities rather than movement alone [16].

Operational stakeholders require different forms of support. Transit operators use demand forecasts,

arrival predictions, disruption alerts, and vehicle-allocation recommendations under time pressure. Emergency services use incident detection, severity estimation, and dynamic routing while retaining authority to respond to conditions absent from the data. Municipal officials and road engineers use predictive evidence to prioritize inspections, redesign streets, adjust signals, and allocate safety investments. Public decision-makers define policy objectives, procurement rules, accountability mechanisms, and acceptable trade-offs among safety, efficiency, accessibility, privacy, and equity. The same prediction consequently requires different explanations, confidence thresholds, and human oversight depending on whether it is addressed to a driver, operator, engineer, or public authority. Transportation-specific AI risk guidance likewise emphasizes the interaction of technical, organizational, and human factors [17]; these differentiated stakeholder needs motivate a framework that links model evaluation to users, decisions, risks, and institutional responsibilities.

4. The Proposed HCT-ML Framework

This section introduces the structure and operational logic of HCT-ML. It first situates the framework relative to established responsible-AI and transportation-assurance instruments, then formalizes its non-compensatory deployment gates, defines measurable evidence for each human-centric dimension, and concludes with context-specific prioritization and explicit trade-off rules.

4.1 Derivation, relationship to existing frameworks, and scope

The six HCT-ML dimensions were selected through a framework-mapping criterion rather than presented as newly discovered ethical principles. A candidate dimension was retained when (i) it is supported by established trustworthy/responsible-AI guidance or transportation assurance literature, and (ii) it directly changes how a transport prediction should be interpreted, authorized, or converted into an intervention. The NIST Artificial Intelligence Risk Management Framework (AI RMF) identifies validity and reliability, safety, security and resilience, accountability and transparency, explainability and interpretability, privacy enhancement, and fairness as trustworthiness characteristics, and organizes risk management through GOVERN, MAP, MEASURE, and MANAGE [12]. The OECD AI Principles similarly emphasize human-centred values and fairness, transparency and explainability, robustness/security/safety, and accountability [18]. The EU AI Act imposes, for relevant high-risk systems, requirements including risk management, data governance, transparency, human oversight, accuracy, robustness, and cybersecurity [19]. IEEE 7000-series standards address value-sensitive system design, measurable transparency, and algorithmic bias [20, 21, 22]. USDOT guidance adds transportation-specific assurance and safety-risk framing [17].

HCT-ML therefore does not compete with these instruments. It is best understood as a transport decision-support profile that can sit inside broader governance or assurance processes. Its organizing unit is not the organization or AI lifecycle as a whole, but a specific prediction-to-decision pathway defined by a decision owner, target population, operational horizon, feasible intervention set, and consequence of error. Table 2 makes this distinction explicit.

Table 2. Benchmark of HCT-ML against established responsible-AI and assurance instruments.

Instrument	Primary organizing scope	Relevant overlap	Relationship to HCT-ML
NIST AI RMF 1.0 [12]	Cross-sector AI risk management across GOVERN, MAP, MEASURE, and MANAGE	Validity/reliability, safety, transparency, explainability, privacy, fairness, accountability, resilience	HCT-ML can function as a transport use-case profile: it narrows broad risk outcomes to a named decision owner, decision horizon, intervention pathway, and dimension-specific gate.

Continued on next page

Table 2 – continued from previous page

Instrument	Primary organizing scope	Relevant overlap	Relationship to HCT-ML
OECD AI Principles [18]	High-level values and policy principles	Human-centred values/fairness, transparency/explainability, robustness/security/safety, accountability	HCT-ML converts selected principles into transport-decision evidence and candidate indicators; it is not a substitute for the broader policy principles.
EU AI Act [19]	Risk-based legal requirements for AI placed on or used in the EU	Risk management, data quality, transparency, human oversight, accuracy, robustness, cybersecurity	HCT-ML is not a legal-compliance framework; it can structure operational evidence concerning intended use, output interpretation, oversight, and fallback within applicable legal obligations.
IEEE 7000/7001/7003 [20, 21, 22]	Engineering processes for ethical concerns, transparency, and algorithmic bias	Value-sensitive design, testable transparency, bias assessment	HCT-ML integrates these concerns at the transportation decision interface and connects them to timing, intervention feasibility, and consequences of error.
USDOT AI assurance guidance [17]	Transportation-specific AI risk and safety assurance	Technical, organizational, and human factors; risk identification and mitigation	HCT-ML adds a compact decision-support protocol for linking prediction evidence to stakeholder action, uncertainty handling, and post-deployment recourse.
HCT-ML (this work)	A specific ITS prediction-to-decision pathway	Relevance, explainability, fairness, uncertainty, human oversight, actionability	Complementary operational profile with measurable evidence, context-specific thresholds, non-compensatory gates, and explicit intervention/fallback mapping.

Several adjacent trustworthy-AI concerns were considered but are not represented as separate HCT-ML dimensions to avoid double counting. Privacy and cybersecurity are treated as cross-cutting legal and engineering constraints that must be satisfied before deployment; they cannot be traded away through a high human-centric score. Robustness and reliability are represented primarily in the technical-performance gate and in uncertainty/distribution-shift assessment. Accountability is treated as a governance outcome implemented through named decision ownership, logging, escalation, recourse, and human oversight. Safety is the target outcome rather than a dimension: HCT-ML specifies plausible decision-support mechanisms that should subsequently be tested against safety-relevant outcomes. This boundary explains why the framework has six dimensions rather than reproducing every category found in NIST, OECD, EU, or IEEE instruments.

4.2 Formal operationalization and non-compensatory deployment gates

Let p denote a task-appropriate predictive-performance measure and τ_p its minimum acceptable threshold. HCT-ML first requires the technical gate $p \geq \tau_p$. For each human-centric dimension $i \in \{R, E, F, U, O, A\}$, let m_{ij} denote an observable indicator, $q_{ij} = g_{ij}(m_{ij}) \in [0, 1]$ a documented normalization or rubric score, and α_{ij} the within-dimension evidence weight. The dimension score is defined in Equation (1):

$$h_i = \sum_{j=1}^{K_i} \alpha_{ij} q_{ij}, \qquad \alpha_{ij} \geq 0, \qquad \sum_{j=1}^{K_i} \alpha_{ij} = 1. \tag{1}$$

The normalization function g_{ij} is not universal: for a calibration error it may decrease as error grows, whereas for user comprehension or action coverage it increases with performance. The protocol

must publish the raw indicator, normalization rule, evidence source, and threshold so that the score is auditable rather than impressionistic.

For safety- or rights-critical dimensions collected in the set C, deployment eligibility is non-compensatory and is defined by Equation (2):

$$G = \mathbb{1}\left[p \geq \tau_p \ \wedge \ h_i \geq \tau_i \ \ \forall i \in C\right]. \tag{2}$$

A system with $G = 0$ is not considered ready for the specified use case regardless of its aggregate score. Only after the required gates pass may stakeholders compute the optional within-use-case synthesis shown in Equation (3):

$$S_{\mathrm{HCT}} = \sum_{i=1}^{6} w_i h_i, \qquad w_i \geq 0, \qquad \sum_{i=1}^{6} w_i = 1, \tag{3}$$

where w_i reflects a documented context-specific priority derived from the decision's risk, urgency, affected population, and available fallback. Equation (3) is therefore a secondary deliberative summary, not a validated universal index and not a basis for ranking unrelated systems. This resolves the earlier ambiguity in which a weighted sum was presented without an explicit measurement protocol or deployment rule.

Figure 1 summarizes the resulting framework and shows how the six human-centric dimensions extend predictive evaluation toward operational decision support.

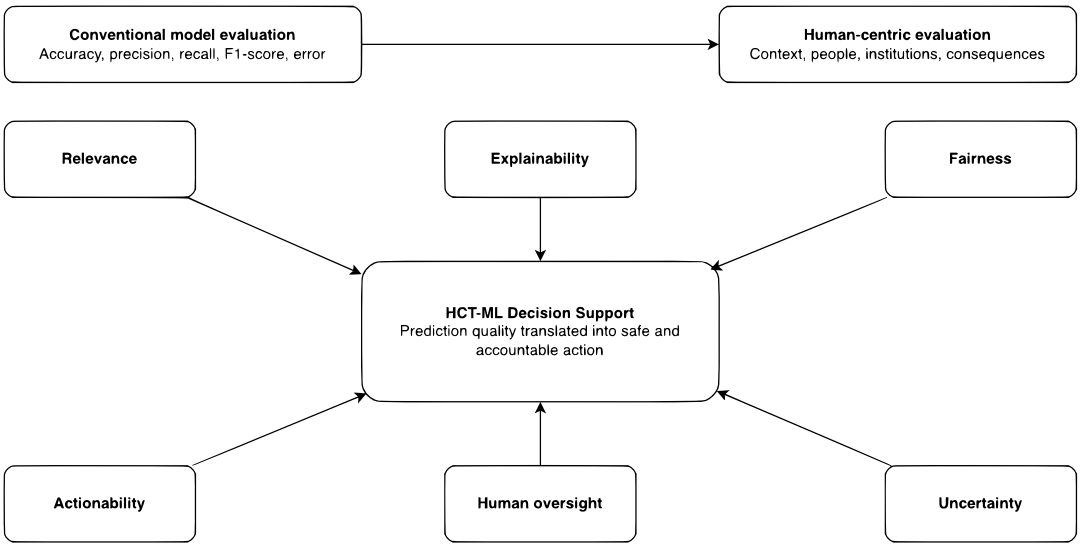

Figure 1. The HCT-ML framework extends conventional predictive evaluation through six dimensions that connect model performance to human-centric decision support

4.3 Dimension definitions, measurable evidence, and safety pathways

Relevance is the fit between a model output and a defined transport decision. Explainability concerns whether the intended user can understand the basis, limits, and practical meaning of an output [7, 8]. Fairness concerns how errors, benefits, and burdens are distributed across people, modes, and places [10, 11, 23]. Uncertainty concerns whether confidence is calibrated and whether unfamiliar

or weakly represented out-of-distribution (OOD) conditions are recognized; calibration measures such as expected calibration error or Brier-type scoring can complement conventional discrimination metrics [24, 25]. Human oversight concerns whether competent people have the authority, information, and time to monitor, question, override, or stop an automated recommendation [26, 19]. Actionability concerns whether a prediction can be translated into a feasible, proportionate, and timely intervention rather than remaining an informational output.

Table 3 operationalizes these definitions. The listed indicators are candidate measures, not mandatory universal metrics. Each deployment should pre-register which indicators are applicable and specify thresholds before evaluating the system. For example, an action-coverage indicator can be defined as $C_A = N_{\text{feasible}}/N_{\text{decision-relevant}}$, while an oversight-coverage indicator can be defined as $C_O = N_{\text{high-consequence outputs reviewed}}/N_{\text{high-consequence outputs}}$. Such indicators turn the dimensions into auditable evidence without claiming that the same numerical threshold is appropriate for every transport mode.

Table 3. Operational protocol for the six HCT-ML dimensions.

Dimension	Evidence protocol	Candidate indicators / gate examples	Proposed safety mechanism
Relevance	Document decision owner, intended use, target population, decision horizon, feasible interventions, and false-positive/false-negative costs before model selection.	Decision-coverage rate; task-success rate; proportion of outputs arriving inside the usable decision window. Gate example: every deployed output class maps to a named owner and intervention pathway.	Prevents technically valid outputs from triggering irrelevant, mistimed, or institutionally infeasible actions.
Explainability	Test explanations with representative users performing the actual decision task, not only with developers.	User comprehension accuracy; time-to-correct interpretation; inappropriate-reliance rate; explanation stability. Gate: predefined comprehension level without material degradation of response time.	Reduces blind reliance, rejection of useful advice, and misunderstanding of system limitations.
Fairness	Report disaggregated errors, calibration, coverage, and benefit/burden allocation across relevant demographic, geographic, modal, and vulnerability groups.	False-negative-rate gap; calibration gap; geographic coverage disparity; worst-group performance. Gate: maximum disparity or minimum worst-group performance defined by policy and data uncertainty.	Reduces systematic under-detection, under-service, or disproportionate exposure to harms for specific groups or areas.
Uncertainty	Evaluate calibration, distribution shift, out-of-distribution conditions, and confidence-triggered fallback.	Expected calibration error; Brier score; risk–coverage curve; OOD detection rate; frequency of unsupported high-confidence outputs. Gate: calibrated confidence plus a tested fallback for low-confidence/OOD states.	Prevents uncertain predictions from being treated as certain and links uncertainty to conservative action or review.
Human oversight	Define authority, competence, monitoring duties, override/stop mechanisms, escalation, logging, and training; test under realistic time pressure.	Oversight coverage C_O; intervention latency; successful override rate; unresolved alert rate; automation-reliance errors. Gate: high-consequence actions are reviewable or interruptible within the decision horizon.	Preserves responsibility and provides recovery when models or operating conditions depart from assumptions.

Continued on next page

Table 3 – continued from previous page

Dimension	Evidence protocol	Candidate indicators / gate examples	Proposed safety mechanism
Actionability	Map each decision-relevant output to a feasible intervention, responsible unit, resource constraint, legal constraint, and response time.	Action coverage C_A; decision-to-action latency; feasibility rate; proportion of recommendations with a documented fallback. Gate: minimum action coverage and response latency below the operational window.	Ensures predictions can produce timely, proportionate safety interventions rather than merely risk scores.

4.4 Context-specific prioritization and trade-offs

The six dimensions are not assumed to have equal importance. Their thresholds and weights depend on the transportation decision. Table 4 provides illustrative priority profiles derived from the dominant failure modes of each application. These profiles are hypotheses for deployment design, not empirically validated rankings.

Table 4. Illustrative HCT-ML priority profiles for transportation applications.

Application	Dimensions that should normally receive the strictest gates	Rationale	Typical evidence
Automated driving / collision avoidance	Uncertainty, human oversight, actionability; explainability for operators and safety assessors	Decisions are time-critical, distribution shift can be abrupt, and fallback time may be short.	Calibration/OOD tests, intervention latency, driver/remote-operator monitoring, tested minimum-risk fallback.
Municipal infrastructure investment	Fairness, relevance, human oversight, explainability	Model outputs can redistribute long-term public resources and burdens across neighbourhoods and road-user groups.	Disaggregated geographic errors, coverage audits, decision rationale, public/engineering review, documented intervention alternatives.
Real-time traffic management	Actionability, uncertainty, relevance	Forecast value decays quickly and interventions must be feasible within current signal/network constraints.	Decision-to-action latency, calibration under incidents, action coverage, operator task performance.
Emergency routing / incident response	Relevance, uncertainty, human oversight, actionability	Local conditions may change faster than the data, and responders require authority to override recommendations.	Route-confidence/fallback logic, override tests, communications latency, blocked-route exception handling.

Trade-offs should be documented as explicit decision choices rather than hidden in one score. NIST notes that trustworthiness characteristics can conflict and that the appropriate balance depends on context [12]. In HCT-ML, fairness versus aggregate predictive performance should therefore be reported through disaggregated performance and, where relevant, constrained or Pareto-style comparisons rather than by silently sacrificing a vulnerable group for a small global gain. Explainability versus model complexity should be evaluated at the user-task level: a more complex model is acceptable only if its explanations and limitations remain usable for the intended decision. Automation versus human oversight is likewise a function-allocation problem [26]: additional review can increase latency, while insufficient oversight can encourage overreliance. The resolution rule is to

identify which dimension is safety- or rights-critical for the use case, impose its minimum gate first, and optimize secondary objectives only within the feasible region created by those gates.

5. Worked Application to Urban Road Safety

To demonstrate implementation without overstating validation, HCT-ML is applied as a worked protocol to a municipal collision-severity decision-support scenario in Montréal. The city publishes police-reported collision data and pursues a Vision Zero road-safety strategy [27, 28]. The scenario uses a real institutional and data context, but no new predictive model, stakeholder experiment, or causal safety evaluation is reported. Its purpose is to show exactly what evidence would be required before a municipal model could be considered deployment-ready under HCT-ML.

Figure 2 summarizes the closed municipal decision cycle through which prediction, human validation, intervention, monitoring, and model revision are connected.

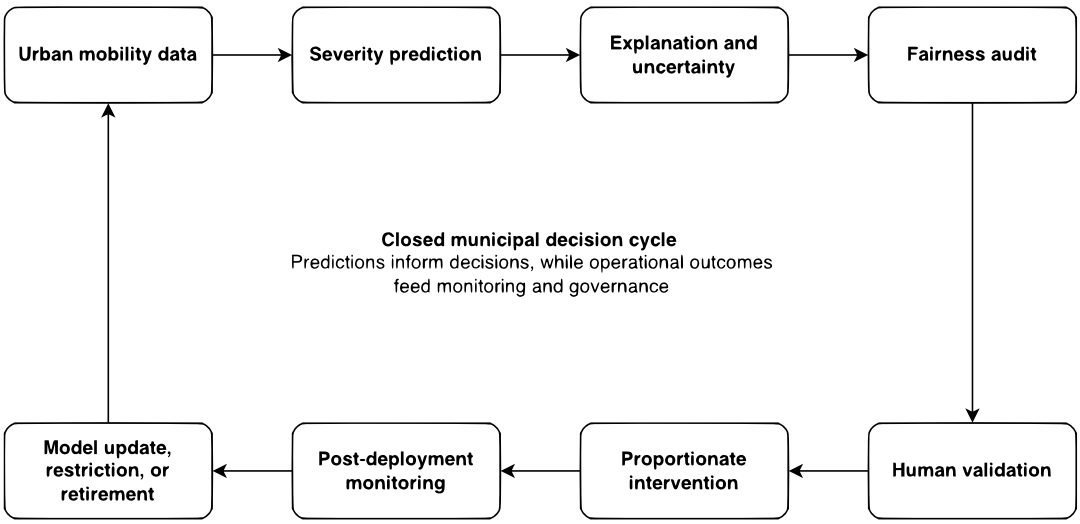

Figure 2. Illustrative application of HCT-ML to a municipal collision-severity prediction system. The workflow is conceptual and does not represent an empirical study

Assume that a municipality develops a model combining historical collision records with road configuration, surface condition, weather, lighting, traffic exposure, time of day, and the presence of vulnerable road users. The decision owner is not "the city" in the abstract but a specified municipal road-safety unit. The intended decision is to prioritize locations for field inspection or engineering assessment within a defined planning window; the model is not authorized to select infrastructure investments automatically. This definition establishes the relevance gate before any comparison of model architectures.

The user interface must then expose the factors and limitations that matter to that decision. A high predicted severity associated with poor lighting, turning complexity, winter conditions, or pedestrian exposure should be presented as predictive association rather than causal proof. Confidence information should accompany each estimate, and locations with sparse collision history, incomplete sensing, or unfamiliar conditions should be routed to additional review instead of receiving a definitive rank. Fairness assessment should compare errors and data coverage across boroughs, road types, and relevant road-user groups, including pedestrians, cyclists, older adults, and people with reduced mobility. Human oversight requires a named engineer or safety officer with authority to

reject a ranking when site knowledge, construction, temporary closure, or other unmodelled conditions invalidate it. Actionability requires every prioritized location to map to a feasible next step, responsible unit, and response window.

Table 5 translates these requirements into the evidence to be collected, illustrative deployment gates, and fallback actions for the Montréal use case.

Table 5. Worked HCT-ML assessment protocol for a Montréal municipal collision-severity tool.

Dimension	Evidence to collect before deployment	Illustrative gate	Action if the gate fails
Relevance	Named municipal decision owner; target road unit; planning horizon; allowed intervention classes; error-cost statement.	No model output is deployed unless it maps to a defined inspection/assessment decision.	Redefine target/output or do not deploy for that decision.
Explainability	Representative municipal users interpret sample cases and identify output limitations.	User-comprehension and task-time thresholds are met in scenario testing.	Redesign explanation/interface or use a more interpretable model for the decision.
Fairness	Borough-, road-type-, and road-user-group error/coverage audit with uncertainty intervals.	No predefined group falls below the municipality's minimum acceptable performance/coverage threshold without mitigation.	Collect data, adjust model/decision rule, limit scope, or require manual review for affected groups/areas.
Uncertainty	Calibration and distribution-shift testing across seasons, weather, and sparse-data locations.	Low-confidence/OOD conditions trigger review rather than definitive ranking.	Recalibrate, restrict use, or activate fallback/manual assessment.
Human oversight	Role assignment, override test, logging, escalation, and review of conflicting site knowledge.	Authorized staff can review/override within the planning window and overrides are auditable.	Suspend automated ranking for affected cases until oversight is functional.
Actionability	Mapping from each risk category to feasible inspection, temporary measure, or engineering assessment; resource/time check.	Required proportion of flagged cases has a feasible action within the defined response window.	Change output granularity, revise thresholds, or treat prediction as research-only information.

After deployment, the same profile becomes a monitoring protocol. The municipality would track calibration drift, mobility and reporting changes, geographic disparities, override frequency, intervention completion, and cases in which predictions repeatedly fail to yield feasible actions. Governance rules should permit the model to be updated, restricted, or suspended when its assumptions no longer hold. This worked application demonstrates operational completeness—what must be measured, who must act, and what happens when a requirement fails—but it should not be interpreted as empirical evidence that HCT-ML improves collision outcomes. Establishing such an effect requires prospective or quasi-experimental evaluation.

6. Research, Validation, and Deployment Agenda

The next research step is empirical validation of the operational protocol rather than further expansion of the dimension list. Validation should occur at three levels. First, measurement validity should test whether the proposed indicators reliably capture the intended constructs. For example, explanation quality should be assessed with intended users performing transport decisions [7], uncertainty measures should be evaluated for calibration and distribution shift [24, 25], and fairness

reporting should remain disaggregated across relevant groups and geographies [10, 11]. Inter-rater agreement should be reported when rubric-based evidence is used.

Second, decision validity should test whether applying HCT-ML changes decision quality relative to accuracy-only evaluation or an established baseline governance process. Suitable study designs include controlled scenario experiments with operators, before/after simulation studies, and prospective field pilots. Outcomes should include decision accuracy, decision latency, appropriate reliance, override quality, subgroup impacts, action completion, and workload rather than only model accuracy. A comparative study could explicitly contrast (a) model-performance-only deployment, (b) NIST/organizational risk-management practice without a transport-specific profile, and (c) the HCT-ML profile layered on the same technical model.

Third, safety and public-value validation should examine downstream outcomes without assuming causality from the conceptual framework. Depending on the application, these may include conflict rates, near misses, incident-response time, intervention effectiveness, accessibility, distribution of safety investment, or calibrated safety surrogates. Prospective designs should pre-specify confounders and distinguish predictive association from causal intervention effects. The title and conclusions of this article therefore avoid claiming that HCT-ML has already produced measurable safety gains.

Operational evaluation must extend beyond retrospective datasets. Transit operators, emergency personnel, planners, engineers, and affected communities should participate in scenario-based tests, simulations, and controlled field trials. Documentation should state intended use, data provenance, known limitations, excluded populations, calibration, uncertainty, conditions that degrade performance, and consequences of false alarms and missed events. These practices align with broader lifecycle risk-management and transportation-assurance guidance [12, 17].

Deployment should begin, not end, evaluation. Monitoring, escalation, and retirement criteria should be treated as lifecycle requirements. Results should remain disaggregated so that disparities are not hidden by stable averages. Human recourse must remain available: operators should be able to question, override, and escalate recommendations, while affected communities need channels for reporting recurring errors or harms. The practical research objective is therefore to determine whether the explicit gates and evidence protocols introduced here improve decision quality, appropriate reliance, and intervention quality compared with simpler accuracy-centred practice.

7. Conclusion

This Perspective has reframed HCT-ML as an operational transport decision-support profile rather than a claim to a new set of ethical principles or a validated universal score. Its six dimensions—relevance, explainability, fairness, uncertainty, human oversight, and actionability—are derived from established trustworthy-AI and transportation-assurance concerns but organized around a specific prediction-to-decision pathway. The principal contribution is the combination of measurable evidence protocols, context-specific thresholds, non-compensatory deployment gates, and explicit mappings from model output to decision owner, intervention, fallback, and affected population.

The comparison with NIST, OECD, EU, IEEE, and USDOT instruments clarifies that HCT-ML is complementary: broader frameworks govern organizational and lifecycle risk, whereas HCT-ML provides a compact profile for the operational decision interface. Real-world automated-driving investigations illustrate why system safety depends on oversight, communicated limitations, and fallback design in addition to technical capability, while the Montréal example shows how the protocol can be instantiated without claiming empirical validation.

HCT-ML should now be tested, not assumed. Future studies should validate the dimension mea-

sures with practitioners, compare HCT-ML-assisted decisions against accuracy-only and existing-governance baselines, and evaluate downstream safety or public-value outcomes using appropriate experimental or quasi-experimental designs. Until such evidence exists, the framework should be used as a structured evaluation and documentation protocol rather than as proof that a system is safe or as a universal ranking instrument.

Author contributions

Bappa Muktar: Conceptualization, Methodology, Literature Review, Visualization, Writing (Original Draft), and Writing (Review and Editing).

Funding statement

No funding was received to assist with the preparation of this manuscript.

Competing interests

The author declares no relevant financial or non-financial competing interests.

Ethics statement

Not applicable. This Perspective does not report research involving human participants, human data, or animals.

Open data statement

No new data were generated or analysed for this article. The Montréal road-collision dataset referenced in the illustrative application is publicly available through the Ville de Montréal open-data portal, as cited in the manuscript.

Reproducibility statement

This article presents a conceptual perspective and does not report an empirical experiment, software implementation, or newly generated dataset. The proposed HCT-ML framework, evaluation dimensions, equations, tables, and conceptual workflows are fully described in the manuscript. No source code is required to reproduce the conceptual contribution.

References

[1] Mohamed Elassy, Mohammed Al-Hattab, Maen Takruri, and Sufian Badawi. "Intelligent transportation systems for sustainable smart cities". In: *Transportation Engineering* 16 (2024), p. 100252. DOI: 10.1016/j.treng.2024.100252.

[2] Nandkumar Niture and Iheb Abdellatif. "A systematic review of factors, data sources, and prediction techniques for earlier prediction of traffic collision using AI and machine learning". In: *Multimedia Tools and Applications* 84 (2025), pp. 19009–19037. DOI: 10.1007/s11042-024-19599-6.

[3] Asad Abdi and Chintan Amrit. "A review of travel and arrival-time prediction methods on road networks: Classification, challenges and opportunities". In: *PeerJ Computer Science* 7 (2021), e689. DOI: 10.7717/peerj-cs.689.

[4] Maged S. AL-Quraishi, Syed Saad Azhar Ali, Muhammad AL-Qurishi, Tong Boon Tang, and Sami Elferik. "Technologies for detecting and monitoring drivers' states: A systematic review". In: *Heliyon* 10.20 (2024), e39592. DOI: 10.1016/j.heliyon.2024.e39592.

[5] Mohammad Hassan Mobini Seraji, Sami Shaffiee Haghshenas, Sina Shaffiee Haghshenas, Vladimir Simic, Dragan Pamucar, Giuseppe Guido, and Vittorio Astarita. "A state-of-the-art review on machine learning techniques for driving behavior analysis: Clustering and classification approaches". In: *Complex & Intelligent Systems* 11 (2025), p. 386. DOI: 10.1007/s40747-025-01988-5.

[6] Ben Shneiderman. "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy". In: *International Journal of Human–Computer Interaction* 36.6 (2020), pp. 495–504. DOI: 10.1080/10447318.2020.1741118.

[7] Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. "Towards human-centered explainable AI: A survey of user studies for model explanations". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.4 (2024), pp. 2104–2122. DOI: 10.1109/TPAMI.2023.3331846.

[8] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. "Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions". In: *IEEE Access* 12 (2024), pp. 101603–101625. DOI: 10.1109/ACCESS.2024.3431437.

[9] Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. "SHIFT: A Synthetic Driving Dataset for Continuous Multi-Task Domain Adaptation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 21371–21382. DOI: 10.1109/CVPR52688.2022.02068.

[10] Yunhan Zheng, Shenhao Wang, and Jinhua Zhao. "Equality of opportunity in travel behavior prediction with deep neural networks and discrete choice models". In: *Transportation Research Part C: Emerging Technologies* 132 (2021), p. 103410. DOI: 10.1016/j.trc.2021.103410.

[11] Xiaojian Zhang, Qian Ke, and Xilei Zhao. "Travel demand forecasting: A fair AI approach". In: *IEEE Transactions on Intelligent Transportation Systems* 25.10 (2024), pp. 14611–14627. DOI: 10.1109/TITS.2024.3395061.

[12] Elham Tabassi. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Tech. rep. NIST AI 100-1. Gaithersburg, MD: National Institute of Standards and Technology, 2023. DOI: 10.6028/NIST.AI.100-1.

[13] National Transportation Safety Board. *Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018*. Tech. rep. NTSB/HAR-19/03. Washington, DC: National Transportation Safety Board, 2019. URL: https://www.ntsb.gov/investigations/Pages/HWY18MH010.aspx.

[14] National Transportation Safety Board. *Collision Between a Sport Utility Vehicle Operating with Partial Driving Automation and a Crash Attenuator, Mountain View, California, March 23, 2018*. Tech. rep. NTSB/HAR-20/01. Washington, DC: National Transportation Safety Board, 2020. URL: https://www.ntsb.gov/investigations/Pages/HWY18FH011.aspx.

[15] World Health Organization. *Global Status Report on Road Safety 2023*. Tech. rep. ISBN 978-92-4-008651-7. Geneva: World Health Organization, 2023. URL: https://iris.who.int/server/api/core/bitstreams/46275f9f-ef66-4892-8ddd-a496ef8c1b74/content (visited on 09/05/2026).

[16] OECD/ITF. *Sustainable Accessibility for All*. Tech. rep. Paris: OECD Publishing, 2024. DOI: 10.1787/5c91857c-en.

[17] Huafeng Yu, Daniel Hulse, and Lukman Irshad. *Understanding AI Risks in Transportation: AI Assurance for Transportation Whitepaper Series*. Tech. rep. Washington, DC: U.S. Department of Transportation, Highly Automated Systems Safety Center of Excellence, Sept. 2024. URL: https://www.transportation.gov/sites/dot.gov/files/2024-09/HASS_COE_AI_Assurance_Whitepaper_AI_Risk_Sep2024.pdf.

[18] Organisation for Economic Co-operation and Development. *OECD AI Principles*. Updated May 2024. 2024. URL: https://oecd.ai/en/ai-principles.

[19] European Parliament and Council of the European Union. *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

[20] IEEE Standards Association. *IEEE 7000-2021: Standard Model Process for Addressing Ethical Concerns during System Design*. 2021. URL: https://standards.ieee.org/standard/7000-2021.html.

[21] IEEE Standards Association. *IEEE 7001-2021: Standard for Transparency of Autonomous Systems*. 2021. URL: https://standards.ieee.org/ieee/7001/6929/.

[22] IEEE Standards Association. *IEEE 7003-2024: Standard for Algorithmic Bias Considerations*. 2024. URL: https://standards.ieee.org/ieee/7003/11357/.

[23] Moritz Hardt, Eric Price, and Nati Srebro. "Equality of Opportunity in Supervised Learning". In: *Advances in Neural Information Processing Systems*. Vol. 29. 2016, pp. 3315–3323.

[24] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. "On Calibration of Modern Neural Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Proceedings of Machine Learning Research. 2017, pp. 1321–1330. URL: https://proceedings.mlr.press/v70/guo17a.html.

[25] Alex Kendall and Yarin Gal. "What uncertainties do we need in Bayesian deep learning for computer vision?" In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, pp. 5574–5584.

[26] Raja Parasuraman, Thomas B. Sheridan, and Christopher D. Wickens. "A model for types and levels of human interaction with automation". In: *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30.3 (2000), pp. 286–297. DOI: 10.1109/3468.844354.

[27] Ville de Montréal. *Collisions routières*. Données ouvertes de la Ville de Montréal. URL: https://donnees.montreal.ca/dataset/collisions-routieres (visited on 07/14/2026).

[28] Ville de Montréal. *Vision Zero: Getting Around Safely on Foot, by Bike and by Car*. Updated 16 May 2025; accessed 14 July 2026. 2025. URL: https://montreal.ca/en/articles/vision-zero-getting-around-safely-foot-bike-and-car-14584.