PHD CANDIDACY PROPOSAL

Reliable Human–AI Decision Support under Distribution Shift

Detecting risk, explaining reliance, and designing safer interventions

Maya Chen
Department of Computer Science • Northbridge University
Advisor: Prof. Elena Hartmann • 18 September 2026

Proposal in one sentence

COMMITTEE DECISION

Can joint monitoring prevent silent failure in AI-assisted decisions?

Decision requested: approve a theoretically grounded, testable, and feasible program.

Important
High-stakes systems drift after deployment.

Testable
Each claim has a falsification criterion.

Feasible
Data, compute, and staged gates are identified.

PART 1

Problem and thesis

The reliability gap appears where changing data meets changing human behavior.

Silent failure sustains trust

OPERATIONAL PATTERN

1. A model is validated on historical data.
2. Deployment changes the data and workflow.
3. Accuracy erodes before labels arrive.
4. Users cannot see when reliance is unsafe.

CORE PROBLEM

Current monitoring asks whether the *model* shifted. Safe deployment requires knowing whether the **joint human–AI decision** became unreliable.

Scope: decision-support settings with consequential outcomes, delayed labels, and a human who can accept, override, or ignore an AI recommendation.

The gap spans three fields

SHIFT DETECTION

Measures changes in inputs, outputs, or representations.

Missing: downstream human response.

HUMAN RELIANCE

Studies trust, calibration, explanations, and overrides.

Missing: changing deployment data.

ADAPTIVE SYSTEMS

Updates models using retraining or test-time adaptation.

Missing: behavioral safety objectives.

Research gap: no unified account links detectable shift, changing reliance, and an intervention that improves the final decision.

One thesis and three questions

Thesis. Deployment risk can be identified earlier and reduced more reliably when monitoring combines model uncertainty with human reliance signals, then triggers an intervention matched to the observed failure mode.

RQ1 Which unlabeled signals predict degradation in joint decision quality?

RQ2 How does distribution shift causally change acceptance and override behavior?

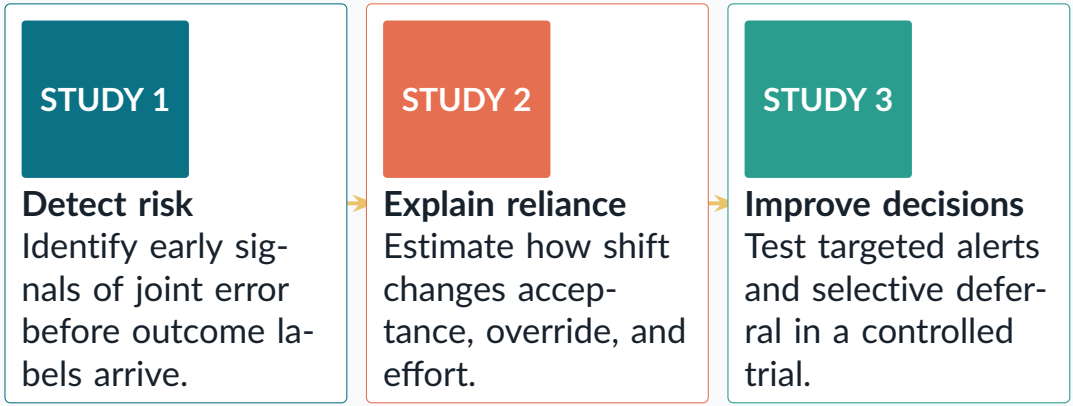
RQ3 Which targeted intervention restores calibration with the least added burden?

PART 2

Proposed research program

Three studies turn one thesis into measurable, falsifiable claims.

Three studies, one argument



Output: a validated monitoring-and-intervention policy for human–AI teams.

Joint decisions are the target

PRIMARY OUTCOME

Decision loss combines incorrect action and task-specific harm.

KEY MODERATOR

User expertise and decision stakes shape the interaction.

Neither uncertainty nor reliance alone determines whether an intervention helps.

Study 1: Detect risk early

STUDY 1

Prospective monitoring study

DESIGN

Replay timestamped cases in order; expose the system to gradual and abrupt shifts; delay outcome labels to mirror deployment.

PRIMARY TEST

Compare signal combinations on lead time and precision for detecting a clinically or operationally meaningful rise in joint error.

SIGNALS

Predictive entropy, representation drift, disagreement, acceptance rate, and response time.

FALSIFICATION

Reject H1 if behavioral signals add no out-of-sample value beyond model-only monitoring.

Study 2: Explain reliance

STUDY 2

Controlled mechanism experiment

MANIPULATION

Randomize shift visibility, model confidence, and explanation quality in a preregistered $2 \times 2 \times 2$ design.

CAUSAL ESTIMAND

Average treatment effect of visible shift cues on appropriate reliance, with expertise as a prespecified moderator.

MEASURED BEHAVIOR

Acceptance, appropriate override, decision time, confidence, and perceived workload.

FALSIFICATION

Reject H2 if shift cues change confidence but do not improve appropriate reliance.

Study 3: Target intervention

STUDY 3 Randomized intervention trial

POLICY ARMS

1. Standard confidence display
2. Contextual shift alert
3. Selective deferral to human review

SUCCESS CRITERION

Lower joint decision loss without a practically unacceptable increase in time or workload.

FALSIFICATION

Reject H3 if the best policy fails the preregistered benefit–burden threshold.

Trigger the intervention by failure mode, not a generic uncertainty threshold.

PART 3

Evidence and execution

Validity comes from converging domains, preregistered tests, and staged decisions.

Two domains test generality

DOMAIN A: CLINICAL TRIAGE

Consequential decisions, delayed outcomes, specialist users, and strong governance constraints.

Role in thesis: high-stakes validation and expert-reliance analysis.

DOMAIN B: INFRASTRUCTURE INCIDENTS

Rapid decisions, rich event logs, recurring operational shifts, and measurable response cost.

Role in thesis: replication, stress testing, and external validity.

Transfer claim: the monitoring logic should transfer; domain-specific harm weights and intervention thresholds should not.

Three claims need three tests

Claim	Primary evidence	Robustness check
Early detection	Lead time at fixed false-alert rate	Time-split and domain-shifted test sets
Causal mechanism	Preregistered treatment effects	Expertise and task-stakes interaction
Practical utility	Joint decision loss	Workload, delay, and subgroup performance

Analysis discipline: one confirmatory model per primary claim; exploratory models are labeled and reported separately.

Feasibility is gated

AVAILABLE NOW

- Data-use pathway identified
- Replay pipeline implemented
- Pilot task and measures drafted
- Compute fits existing allocation

GATE BEFORE EACH STUDY

G1: Data quality supports the target shift.

G2: Pilot establishes reliable behavior measures.

G3: Intervention passes usability and safety review.

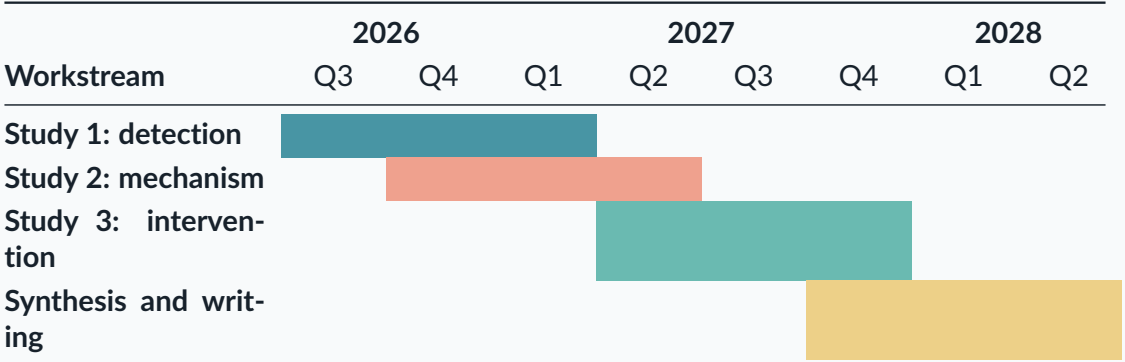
If a gate fails, the study narrows its claim before new data collection begins.

Each risk has a fallback

Risk	Prevention	Fallback
Restricted data	De-identification, minimum access, audit log	Public benchmark plus synthetic shift
Weak behavior signal	Validated task, pilot reliability threshold	Narrow claim to observable actions
Alert fatigue	Burden cap and human-factors review	Selective deferral only
Unequal benefit	Prespecified subgroup analysis	Group-conditional threshold or no deployment

No live clinical recommendation is made during the dissertation; all intervention studies use retrospective replay or controlled simulation until separate deployment approval.

A 24-month critical path



Submission sequence: detection → mechanism → intervention + dissertation.

Dissertation contributions

CONCEPTUAL

A model of how distribution shift, model uncertainty, and human reliance jointly produce risk.

METHODOLOGICAL

Evaluation methods that treat the joint decision, not standalone model accuracy, as the target.

PRACTICAL

A policy for choosing among monitoring, alerting, and deferral under explicit burden constraints.

Dissertation claim: reliable decision support is a property of the adaptive human–AI system, not of the model in isolation.

DECISION REQUEST

Approve a testable program for safer human–AI decisions

The proposal connects early detection, causal explanation, and targeted intervention through explicit evidence gates.

Questions and discussion
Maya Chen • maya.chen@northbridge.edu

Selected references

HUMAN–AI INTERACTION AND UNCERTAINTY

Amershi, S. et al. (2019).
Guidelines for human–AI interaction. *Proceedings of CHI*.

Ovadia, Y. et al. (2019).
Can you trust your model's uncertainty? *Advances in Neural Information Processing Systems*.

Sample references – replace with the literature that directly motivates your proposal.

Selected references, continued

ACCOUNTABILITY AND DEPLOYED SYSTEMS

Raji, I. D. et al. (2020).
Closing the AI accountability gap. *Proceedings of FAccT*.

Sculley, D. et al. (2015).
Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*.

Candidacy criteria

Criterion	Evidence in proposal	Remaining committee input
Significance	Joint reliability affects consequential decisions	Is the scope important enough?
Originality	Connects detection, reliance, and intervention	Is the thesis sufficiently distinct?
Rigor	Preregistered tests and explicit falsification	Are outcomes and estimands adequate?
Feasibility	Staged gates and bounded fallback paths	Are resources and timing realistic?

Committee feedback becomes a revised scope and milestone agreement.