ALTAY CONFERENCE PROCEEDINGS IN MATHEMATICS
2 (2025), No. 1, pp. 15-22
https://altayconfproceedings.com/
ISSN 3108 - 5288

Research Article

Mathematical modelling of control-loss detection via risk-sensitive reinforcement learning on partially observable Markov decision processes

OLEKSANDR CHABAN* AND VOLODYMYR HLADUN

ABSTRACT. This work introduces a method for identifying high-risk loss-of-control episodes in digital settings by combining risk-sensitive reinforcement learning with decision-making under partial observability. We motivate the need to reason with incomplete and noisy information — typical of real-world deployments and, in particular, of monitoring user behaviour during critical states. The agent-environment interaction is modelled within the partially observable Markov decision process formalism, which maintains a belief (probabilistic posterior) over latent states given histories of actions and observations. Behaviour is analysed at the trajectory level, and tail risk is quantified via the Conditional Value at Risk (CVaR), enabling the assessment of expected losses in worst-case regimes rather than average-case performance. To ensure transparency and foster trust, we integrate explainable AI (XAI) techniques that reveal the factors driving risk estimates and action choices. The resulting pipeline provides a principled basis for adaptive detection of critical states and for early-warning interventions in complex digital environments, supporting reliable and accountable decision support.

Keywords: Reinforcement learning, partially observable Markov decision process, risk-sensitive metric, explainable AI techniques.

2020 Mathematics Subject Classification: 90C40, 93C41, 68T05.

1. INTRODUCTION

In rapidly evolving sociotechnical systems, where signals are noisy, structure is complex, and uncertainty is pervasive, there is a growing need to identify hazardous or critical modes in a timely manner. Central to this agenda is the notion of loss of control: episodes in which a human operator or an automated system deviates from intended behavior under incomplete knowledge of the true environmental state. Such episodes can precipitate severe outcomes – ranging from user dependencies and anomalous activity to security-policy violations and high-stakes decision errors.

Models that presuppose full state observability – e.g., classical Markov Decision Processes (MDPs) – are ill-suited to these settings, which are inherently stochastic and only partially observed. A more faithful representation is provided by the Partially Observable Markov Decision Process (POMDP), which maintains a probabilistic belief over latent states updated from the history of actions and observations, thereby offering a principled mechanism for operating under deep uncertainty.

Received: 22.08.2025; Accepted: 06.10.2025; Published Online: 07.11.2025
*Corresponding author: Oleksandr Chaban; oleksandr.m.chaban@lpnu.ua
DOI: 10.64700/altay.17
Presented in *International Workshop on Modern Problems of Analysis, Optimization, Approximation and Their Applications*

A further challenge is explicit treatment of downside risk in atypical or adversarial regimes. Risk-sensitive criteria such as Conditional Value at Risk (CVaR) quantify losses in the distributional tail and, unlike expectation-centric objectives, emphasize the magnitude of adverse outcomes. This focus supports the synthesis of robust, risk-aware policies.

Coupling reinforcement learning with the POMDP formalism yields an adaptive toolkit: the agent incrementally incorporates new evidence, updates its policy from experience, and responds to environmental drift in real time. To promote trust and transparency, explainable AI (XAI) techniques can be used to interpret internal decision pathways and improve model interpretability.

Against this backdrop, the present work introduces a risk-sensitive framework for detecting loss-of-control episodes that integrates POMDP modelling, CVaR-based objectives, and reinforcement learning. The resulting approach enables early warning and identification of critical situations under partial observability and high uncertainty.

2. Illustrative scenarios of control loss

To illustrate the practical scope of the approach, we outline scenarios in which an individual undergoes a transient loss of control – understood here as a behavioral episode that yields direct or indirect adverse outcomes and is later recognized by the individual or their environment as erroneous.

First, consider social-media use. Users may impulsively post messages or comments that are subsequently judged inappropriate or reputation-damaging. Beyond basic content filters, organizations could adopt configurable, policy-aware pre-publication checks – such as a browser extension – that warn when draft content conflicts with internal guidelines. This is especially pertinent for official or corporate accounts.

A second setting is online gambling. During affective or impulsive states, users may assume excessive risk, with substantial financial consequences. As regulation increasingly mandates responsible-gambling safeguards, a natural extension is to integrate predictive modules that detect early indicators of risk-seeking behavior and trigger proportionate interventions (e.g., recommending a break or notifying a pre-designated contact). Selective application to vulnerable cohorts can align user protection with legal and ethical expectations.

A third scenario pertains to software deployment in CI/CD pipelines. Under deadline pressure, engineers may bypass tests or override safeguards, pushing misconfigured builds to production with downstream organizational impact. Predictive modules can monitor contextual and behavioral indicators (e.g., recent failure history, code churn, out-of-hours commits) and trigger proportionate interventions – requesting peer approval, recommending a canary rollout, or deferring deployment to a lower-risk window. These mechanisms operationalize loss-of-control detection in high-stakes technical workflows and align with existing DevOps governance.

These examples ground the abstract notion of control loss in concrete contexts and motivate the value of the proposed modelling framework.

3. The agent as a Markov decision process

The Markov Decision Process (MDP) [2, 9, 15] is adopted as the primary abstraction for agent–environment interaction because it offers a mathematically tractable way to model tasks

in which actions drive stochastic state transitions and immediate outcomes in a dynamic setting. In effect, the MDP encodes sequential decision-making and the causal dependencies between states, actions, and rewards. Formally, an MDP is a tuple

$$\mathcal{M} = \langle S, A, \tau, r, \gamma \rangle,$$

where $\tau(s'|s,a)$ is the transition kernel, $r(s,a) \in \mathbb{R}$ is the one-step reward, and $\gamma \in [0,1]$ is the discount. For any policy $\pi(a|s)$, the value functions satisfy the Bellman relations:

$$V^\pi(s) = \sum_a \pi(a|s) \left[r(s,a) + \gamma \sum_{s'} \tau(s'|s,a) V^\pi(s') \right], Q^\pi(s,a) = r(s,a) + \gamma \sum_{s'} \tau(s'|s,a) \sum_{a'} \pi(a'|s') Q^\pi(s$$

Moreover, the MDP formalism underpins modern Reinforcement Learning (RL) algorithms and theory [18, 19, 20].

A standard MDP presumes full observability: at each time t, the agent observes the true state s_t prior to choosing a_t. In problems concerned with detecting moments of loss of control, this assumption is rarely justified. The monitoring system does not access latent internal conditions (e.g., cognitive or emotional states) and instead receives only indirect, noisy signals. Let z_t denote the observed proxy vector (e.g., reaction time, message or bet size, click frequency, keystroke error rate); decisions must then be made under partial observability.

To explicitly encode uncertainty about the hidden state, we model the problem as a partially observable MDP (POMDP) [11, 13, 17, 18]. The agent maintains a belief state b_t, a probability distribution over latent states representing its posterior information after past actions and observations [4, 16]. Given action a_t and new observation z_{t+1}, the belief is updated by the Bayes filter:

$$b_{t+1}(s') = \eta \, o(z_{t+1}|s', a_t) \sum_{s \in S} \tau(s'|s, a_t) b_t(s), \tag{3.1}$$

with

$$\eta^{-1} = \sum_{s'} o(z_{t+1}|s', a_t) \sum_{s} \tau(s'|s, a_t) b_t(s). \tag{3.2}$$

Planning then proceeds over beliefs, enabling policies that act because of uncertainty rather than in spite of it; empirical performance typically improves in realistic, noisy environments.

From a formal perspective, passing to the belief space $\mathcal{B} = \Delta(\mathcal{S})$ converts the partially observable control problem into a fully observable (generally continuous) belief-MDP. The one–step belief reward and the (stochastic) belief transition induced by the Bayes operator Φ are

$$R(b,a) = \sum_s b(s) R(s,a), \quad b' = \Phi(b,a,z), \tag{3.3}$$

so that the optimal value function solves

$$V^*(b) = \max_{a \in A} \left\{ R(b,a) + \gamma \mathbb{E}_{z \sim p(\cdot|b,a)} \left[V^*(\Phi(b,a,z)) \right] \right\}, \tag{3.4}$$

with $p(z|b,a) = \sum_{s'} o(z|s',a) \sum_s \tau(s'|s,a) b(s)$.

A POMDP can be specified by the tuple

$$\mathcal{P} = \langle S, A, Z, \tau, o, r, \gamma \rangle, \tag{3.5}$$

S – latent state space (true configurations are not directly observable);
A – action space (available controls at each step);

Z – observation space (noisy signals emitted by the system);
$\tau(s'|s,a)$ – state transition kernel;
$o(z|s',a)$ – observation kernel (likelihood of z after arriving in s' via a);
$r(s,a) \in \mathbb{R}$ – immediate reward.

Intuitively, τ captures uncertainty in the underlying dynamics, o connects latent states to observable evidence, and r quantifies instantaneous utility. Decisions are then computed over beliefs $b \in \mathcal{B}$, leveraging the updates above to integrate new information and act optimally under partial information.

The formalism of a partially observable Markov decision process offers a versatile framework for modelling decision-making under uncertainty. To situate its potential, it is helpful to consider uses in neighbouring domains. A broad survey of applications is provided in [3], although the particular use case studied here is not treated explicitly. Structurally closest analogues arise in behavioural ecology (scientific domain), in medical diagnostics (social domain) and in corporate policy (business domain).

In behavioural ecology, a common premise is that an organism behaves (approximately) optimally with respect to an internal representation of its environment. The analyst posits candidate states, actions, observations, and rewards thought to be salient for the organism, derives an optimal strategy from this specification, and compares it with observed behaviour. Mismatches motivate alternative hypotheses and iterative refinement of the model, progressively deepening our understanding of behaviour–environment interactions.

Our focus differs: rather than incrementally perfecting the latent model, we seek to detect marked departures from the putative optimal strategy. Such departures are interpreted as indicators of a loss of self-control in the individual under observation.

In medical diagnostics, despite substantial advances, practice still hinges on expert judgement and remains error-prone. The core difficulty is that the patient's internal state is only partially observable. The clinician must choose among actions–laboratory tests, medication, surgery, physical therapy–each carrying costs and risks, yet providing only partial information about the underlying condition. Consequently, one must balance observational accuracy against financial or medical burden.

In a context relevant to our work, where gambling addiction is formally recognised as a disorder [21, 22], the model's agent can be cast as a virtual "physician" tasked with inferring loss of control from limited, indirect observations. The objective is not merely to explain behaviour post hoc, but to identify early signals of deterioration in a principled, uncertainty-aware manner.

In corporate governance, a strand of POMDP-based work treats organizational control problems – internal audits, accounting, and policy enforcement – as sequential decision making under partial observability [3]. The key premise is system-level: while individual agents (employees) perceive only fragments, an internal model can aggregate telemetry from across the enterprise to infer latent organizational states more reliably than any single decision-maker. This separation between human-limited views and a model's integrated perspective motivates the use of belief-state monitoring for stability and compliance at scale.

Our setting instantiates this perspective in software delivery pipelines. Deployment readiness is only indirectly observable through proxies such as test coverage, recent failure history, code churn, review latency, or out-of-hours activity; actions (e.g., triggering additional tests, requesting peer approval, switching to a canary rollout, or deferring release) trade information gain and risk reduction against cost and delay. Modeling the pipeline as a POMDP allows principled balancing of these trade-offs, and provides a mechanism for detecting loss-of-control

episodes under deadline pressure – when safeguards are bypassed or misconfigurations propagate – by elevating intervention intensity as the belief shifts toward hazardous latent states.

4. Reinforcement Learning as a tool for detecting control loss risk

The central premise of applying reinforcement learning to behavioral modelling is to specify an agent that acquires an optimal policy by interacting with its environment. Formally, this interaction is represented as a sequence of states, actions, and rewards, consistent with the structure of a Markov Decision Process (MDP) or, under partial observability, its POMDP extension.

The goal of the agent is to find a policy $\pi(a|s)$ that maximizes the expected cumulative reward:

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)|\pi\right], \tag{4.6}$$

where $\gamma \in [0, 1]$ is the discount factor that governs the relative weight of future returns.

In behavioral analytics, attention shifts from isolated actions or momentary states to the evaluation of the complete behavioral trajectory:

$$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_T), \tag{4.7}$$

which aggregates the downstream effects of the agent's decisions. This trajectory-level view reveals persistent action patterns that can culminate in adverse outcomes. Rather than flagging single "risky" moves, the analysis considers how conditions accumulate over time and gradually elevate the likelihood of control loss.

Within this perspective, it is natural to specify a subset of critical states $s_{crit} \subseteq S$ in which the probability of loss of control (e.g., progression to anomalous or addictive behavior) rises markedly. Behavior is then interpreted not merely as a reaction to the present state but as the cumulative consequence of prior choices. Consequently, a risk-detection mechanism should avoid judging the current action in isolation and instead condition its assessment on the full historical context encoded in τ.

To assess risk in dynamic or stochastic decision-making, we employ a trajectory-level risk-sensitive criterion. This perspective goes beyond average performance to capture the probability mass of rare but consequential failures, thereby measuring the system's vulnerability to adverse, off-nominal conditions that fall outside typical expectations.

Within this setting, the Conditional Value at Risk (CVaR) [14, 23] at confidence level α, denoted $CVaR_\alpha$, is appropriate. CVaR is the expected loss conditional on exceeding a quantile threshold defined by the Value at Risk (VaR) [5, 10, 12]. Equivalently, $CVaR_\alpha$ computes the mean loss over the worst $100 \cdot (1-\alpha)\%$ of outcomes – the tail of the loss distribution:

$$CVaR_\alpha = \mathbb{E}\left[L|L \geq VaR_\alpha\right], \tag{4.8}$$

where L is a loss-valued random variable. In reinforcement learning, L can be instantiated as the cumulative negative reward incurred along a trajectory, interpreting negative reward as a penalty for unsafe or undesirable behavior that degrades policy performance [18].

Unlike VaR – which merely identifies the α-quantile loss not to be exceeded with probability α – CVaR quantifies tail severity, i.e., the expected magnitude of loss once the system has entered worst-case regimes.

While we adopt $CVaR_\alpha$ as the primary risk functional due to its coherence, tail sensitivity, and convexity – properties that facilitate stable optimization under partial observability – it is not the only admissible choice. Alternatives such as entropic (exponential-utility)

risk, mean–variance criteria, and distributional/quantile losses may be preferable in domain-specific deployments. Our pipeline is risk-functional agnostic: $CVaR_\alpha$ can be replaced by other coherent or utility-based measures without modifying the overall design.

For reinforcement learning or trajectory-level analysis, this specialization takes the form

$$CVaR_\alpha \left(-\sum_{t=0}^{T} r_t \right), \tag{4.9}$$

where r_t denotes the immediate reward at time t and the sum spans the entire trajectory. The leading minus sign enforces the standard loss-as-negative-reward convention used in risk-sensitive optimization. Hence, the metric captures the average loss among trajectories with the lowest (worst) cumulative returns.

We define a trajectory-level risk estimator as

$$\mathcal{R}(\tau) : \tau \mapsto [0, 1], \tag{4.10}$$

using data from offline reinforcement learning (offline RL) [7, 8], is trained to discriminate between trajectories that are safe and those that are risky. This enables early warning of potentially hazardous behavioral patterns and supports risk-aware sorting of trajectories, with an emphasis on highlighting critical states.

Because offline RL relies on a fixed logged dataset, the risk estimator is vulnerable to distributional shift between training data and deployment, which can degrade reliability. In this study we restrict interpretation to in-distribution regimes (i.e., behaviour supported by the logged data) and note that out-of-support trajectories require additional safeguards and validation in real-world evaluations.

By computing risk-sensitive metrics at the trajectory level and selecting appropriate decision thresholds, the system can flag episodes in which the likelihood of control loss becomes non-negligible. A complementary ingredient of the approach is explainable AI (XAI) [1, 6], which is essential for making the RL model's behavior transparent and for establishing trust in its outputs – especially in applications involving anomalous usage or emergent behavioral addictions, where model decisions may directly affect a person's psycho-emotional well-being.

Missed detections (false negatives) are particularly consequential: failing to recognize a loss-of-control event can lead to substantial harm. Moreover, the very presence of a monitoring system can induce a perception of external safety, potentially attenuating users' internal safeguards; if the model then fails to intervene when needed, it not only forfeits its protective role but may amplify the resulting harm due to misplaced reliance.

Symmetrically, false positives carry significant costs for organizations deploying such systems. Over-triggering on benign behavior can frustrate users, reduce engagement, and erode trust – ultimately undermining competitiveness and complicating real-world adoption.

For these reasons, XAI plays a central role: it allows developers, users, and independent auditors to understand why a particular trajectory was flagged as risky or anomalous, improving decisional transparency and enabling calibration to application-specific requirements for sensitivity and specificity. Further gains in interpretability can be obtained with attention-based mechanisms that automatically surface the most informative segments of the input – across time or feature dimensions. Visualizing attention weights (e.g., as heatmaps) clarifies which portions of the trajectory or which state features most influenced the risk estimate, providing deeper insight into the agent's decision logic while remaining consistent with the overall model structure.

5. Conclusion

This work examined a risk-sensitive framework for detecting loss-of-control episodes in digital settings by integrating partially observable models (POMDP), reinforcement learning, and risk-aware objectives. We argued that the POMDP formalism is well suited to decision-making under uncertainty typical of real-world systems in which full state observability is infeasible.

Special emphasis was placed on Conditional Value at Risk (CVaR) as a trajectory-level criterion for downside risk. By quantifying tail losses in worst-case regimes, CVaR introduces a principled layer of caution, steering policies not only toward maximizing expected returns but also toward avoiding low-probability, high-impact failures.

We also justified the use of explainable AI (XAI) to expose internal decision pathways, clarify causal relations, and strengthen trust among end users and operators.

Taken together, the proposed concept supports the development of adaptive intelligent systems capable of: detecting critical shifts in behavioral patterns, providing explanations of decisions, synthesizing control strategies that remain robust under uncertainty and rare but destructive events.

Using both online and offline reinforcement learning yields a flexible foundation for real-time systems that adapt to environmental changes, accumulate experience, reassess risk profiles, and preempt critical outcomes.

While the proposed framework shows promise under controlled conditions, it has not yet been validated on large-scale, real-world datasets; consequently, its practical robustness remains unproven. We therefore interpret the present results as feasibility evidence rather than deployment-ready performance, particularly in the presence of distribution shift and operational constraints.

In future work, we plan to instantiate the framework in concrete domains, such as detecting critical behavioral states on online platforms and analyzing autonomous agents operating under severe uncertainty. We will also pursue advanced XAI mechanisms that provide multistep explanations for sequential decisions. Finally, data-driven structural and parametric identification of POMDP models from empirical evidence – beyond the scope of this study – will be addressed to enable practical deployment and validation.

Acknowledgments

The authors gratefully acknowledge the Armed Forces of Ukraine, whose service and sacrifice ensured the safety and conditions necessary to complete this research.

They also acknowledge the Department of Applied Mathematics at Lviv Polytechnic National University for its support during the preparation of this study and for the trust placed in its research potential.

Author contributions. Conceptualization, V.H.; investigation, O.C.; writing–original draft, O.C; writing–review & editing, V.H.; project administration, V.H. All authors have read and agreed to the published version of the manuscript.

Financial disclosure. None reported.

Conflict of interest. The authors declare no conflicts of interest.

References

[1] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. G. Chatila and F. Herrera: *Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,* Information Fusion, **58** (2020), 82–115.

[2] R. J. Boucherie, N. M. van Dijk: *Markov decision processes in practice,* Springer International Publishing, Cham, Switzerland (2017).
[3] A. R. Cassandra: *A survey of POMDP applications,* Working notes of the AAAI 1998 Fall Symposium on Planning with Partially Observable Markov Decision Processes, (1998), 17–24.
[4] X. Chen, Y. Mu, P. Luo, S. E. Li and J. Chen: *Flow-based recurrent belief state learning for POMDPs,* Proceedings of the 39th International Conference on Machine Learning, **162** (2022), 3444–3468.
[5] D. Duffie, J. Pan: *An overview of value at risk,* Journal of Derivatives, **4** (3) (1997), 7–49.
[6] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan and R. Ranjan: *Explainable AI (XAI): Core ideas, techniques, and solutions,* ACM Computing Surveys, **55** (9) (2023), 1–33.
[7] R. Figueiredo Prudencio, M. R. O. A. Maximo and E. L. Colombini: *A survey on offline reinforcement learning: Taxonomy, review, and open problems,* IEEE Transactions on Neural Networks and Learning Systems, **35** (8) (2024), 10237–10257.
[8] Y. Fu, D. Wu and B. Boulet: *A closer look at offline RL agents,* Advances in Neural Information Processing Systems, **35** (2022), 8591–8604.
[9] F. Garcia, E. Rachelson: *Markov decision processes,* Markov decision processes in artificial intelligence, Wiley, (2013), 1–38.
[10] P. Jorion: *Risk2: Measuring the risk in value at risk,* Financial Analysts Journal, **52** (6) (1996), 47–56.
[11] L. P. Kaelbling, M. L. Littman and A. R. Cassandra: *Planning and acting in partially observable stochastic domains,* Artificial Intelligence, **101** (1-2) (1998), 99–134.
[12] T. J. Linsmeier, N. D. Pearson: *Value at risk,* Financial Analysts Journal, **56** (2) (2000), 47–67.
[13] W. S. Lovejoy: *A survey of algorithmic methods for partially observed Markov decision processes,* Annals of Operations Research, **28** (1) (1991), 47–65.
[14] X. Ni, L. Lai: *Robust risk-sensitive reinforcement learning with conditional value-at-risk,* Proceedings of the 2024 IEEE Information Theory Workshop, (2024), 520–525.
[15] M. L. Puterman: *Markov decision processes*, Stochastic models, Handbooks in Operations Research and Management Science, Elsevier, **2** (1990), 331–434.
[16] N. Roy, G. Gordon and S. Thrun: *Finding approximate POMDP solutions through belief compression,* Journal of Artificial Intelligence Research, **23** (2005), 1–40.
[17] A. S. Sinha, A. Mahajan: *Agent-state based policies in POMDPs: Beyond belief-state MDPs,* Proceedings of the 63rd IEEE Conference on Decision and Control, (2024), 6722–6735.
[18] R. S. Sutton, A. G. Barto: *Reinforcement learning: An introduction,* (1st ed.), Cambridge, MA: MIT Press (1998).
[19] C. Szepesvári: *Reinforcement learning algorithms for MDPs,* Wiley encyclopedia of operations research and management science, (2011).
[20] M. A. Wiering, M. Van Otterlo: *Reinforcement learning,* Adaptation, Learning, and Optimization, **12**, Springer, Berlin, Germany (2012).
[21] *Inclusion of gaming disorder in ICD-11,* World Health Organization (2018). Retrieved August 15, 2025, `https://www.who.int/news/item/14-09-2018-inclusion-of-gaming-disorder-in-icd-11`
[22] *Gaming disorder,* World Health Organization (2025). Retrieved August 15, 2025, `https://www.who.int/standards/classifications/frequently-asked-questions/gaming-disorder`
[23] Y. Zhao, W. Zhan, X. Hu, H. F. Leung, F. Farnia, W. Sun and J. D. Lee: *Provably efficient CVaR RL in low-rank MDPs,* arXiv preprint, (2023), arXiv:2311.11965.

Oleksandr Chaban
Lviv Polytechnic National University
Department of Applied Mathematics
12 Stepan Bandera Str., 79013, Lviv, Ukraine
Email address: `oleksandr.m.chaban@lpnu.ua`

Volodymyr Hladun
Lviv Polytechnic National University
Department of Applied Mathematics
12 Stepan Bandera Str., 79013, Lviv, Ukraine
Email address: `volodymyr.r.hladun@lpnu.ua`