Delphi: A Neuro-Symbolic Framework for Individualized, Safe, and Interpretable Treatment Recommendation

Muchan Tao[1*], Haonan Qin[1*], Yuqi Fang[1†], Caifeng Shan[1†], Tieniu Tan[1†]

[1]School of Intelligence Science and Technology, Nanjing University, Nanjing, China
{602024710012, 221900460}@smail.nju.edu.cn, {yqfang, cfshan, tnt}@nju.edu.cn

Abstract

Clinical reinforcement learning (RL) holds promise for treatment recommendation. However, its adoption is hindered by black box decision processes, limited safety guarantees, and a lack of individualized treatment. To address these issues, we introduce Delphi, the first trainable neuro symbolic causal RL framework for dynamic treatment planning, designed to answer three core clinical questions: Why for this patient? Why is it safe? Why this action? Specifically, Delphi constructs: 1) *causality aware state modeling*, using discretized physiological variables and subgroup specific causal graphs; 2) *adaptive symbolic rule constraints*, combining clinical guidelines and behavior-derived rules into the RL system; and 3) *interpretable decision fusion*, where actions are selected based on joint neural symbolic Q values and explained via structured LLM-based justifications. We evaluate Delphi on MIMIC-III sepsis cohort with more than 20,000 trajectories, and experiments show that our Delphi achieves leading performance among existing methods. Moreover, Delphi introduces the first blinded physician evaluation of an explainable RL system in healthcare. Results demonstrate that Delphi consistently outperforms historical physicians' treatments in six dimensions, including adoption rate (+5.75%), understandability (+8.9%), safety (+10.4%), satisfaction (+9.35%), trust (+8.78%), effectiveness (+8.87%). These results highlight Delphi's potential as an interpretable, safe, and patient-specific AI assistant for critical care medicine.

Introduction

Sepsis, a life-threatening organ dysfunction triggered by infection, affects millions globally and accounts for 20% of all deaths worldwide (Rudd et al. 2020). Treating sepsis is profoundly complex (Singer et al. 2016), requiring timely, precise, and highly individualized interventions due to its pathological heterogeneity and treatments with delayed effects (Rhodes et al. 2017). Identifying the optimal dosage and combination of therapies amidst rapidly evolving patient states remains a major clinical challenge (Fleischmann-Struzek and Rudd 2023).

The need for personalized, sequential treatment planning has led to the adoption of Reinforcement Learning (RL) (Yin et al. 2022; Drudi et al. 2024), which is well-suited for this task due to its ability to optimize long-term outcomes under uncertainty. Prior studies have shown its promising results in the treatments of diseases such as sepsis, cancer, and diabetes (Yin et al. 2022; Drudi et al. 2024; Niraula et al. 2021; Wu et al. 2023). Although reinforcement learning has strong potential, most studies fail to meet clinical standards. They lack clear explanations, have limited safety checks, and cannot adjust to individual patient needs. We argue that interpretability, safety, and personalization are highly critical for clinical use. Specifically, interpretability helps doctors understand and trust the system's recommendations. Safety ensures decisions follow medical guidelines and avoid harm. Personalization makes sure treatments fit each patient's condition and progress. Therefore, to build truly trustworthy clinical AI systems, we must answer: **Why for this patient? Why is it safe? Why this action?**

Why for This Patient at This Time? The foundation of effective clinical RL systems lies in robust patient state representation, which enables the model to answer "why this treatment for this patient at this time?" by accurately capturing each individual's unique and evolving physiological condition. However, current methods are limited. Some approaches simply treat each physiological measurement as a separate data point (Lee et al. 2024; Luo et al. 2024), ignoring the interconnected nature of a patient's symptoms.

Other methods use unsupervised clustering techniques (e.g., k-means++) to represent discrete states (Drudi et al. 2024; Ghasemi et al. 2025), but they may lose patient-specific information and prevent models from dynamically adapting to the evolving pathophysiology of each patient. These constraints may result in one-size-fits-all therapeutic strategies that don't suit the unique circumstances of each individual. To address these issues, we introduce Causality-Aware State Modeling, which dynamically constructs personalized causal graphs that explicitly capture the interdependent nature of a patient's symptoms and physiological variables, thereby ensuring that patient-specific treatment decisions can align with their unique and evolving pathological state.

Why Is It Safe? A reliable AI system should build safety from the start, rather than provide explanations retrofitted afterward (Glanois et al. 2024). Existing recommendation

*These authors contributed equally.
†Corresponding authors.

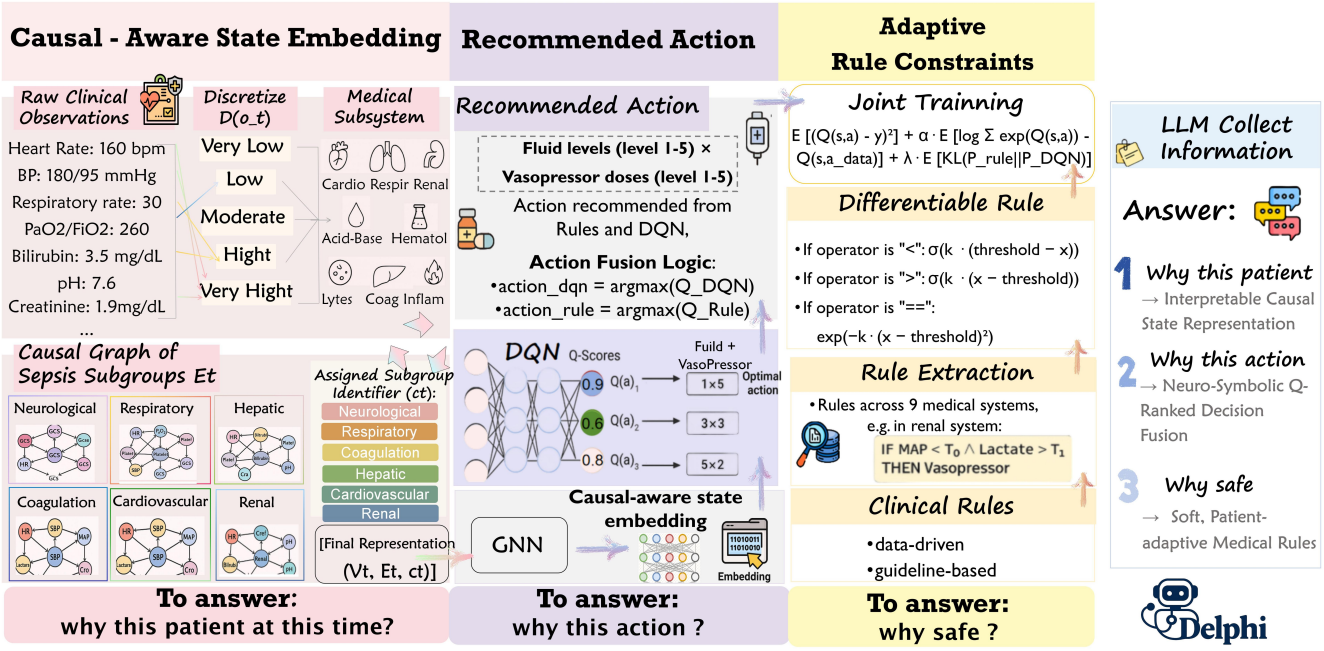

Figure 1: The framework is designed to address three core questions: *Why for this patient at this time?* through interpretable, causality-aware state modeling; *Why is it safe?* through the integration of trainable rules from medical guidelines; and *Why this action?* by LLM-generated explanations. The causality-aware state representation incorporates multiple physiological subsystems, including cardiovascular (Cardio), respiratory, renal, coagulation (Coag), acid–base regulation, electrolyte (Electro), hematologic (Hema), and inflammatory (Inflam) systems, each capturing clinically relevant aspects of organ function and systemic dysregulation.

methods (Komorowski et al. 2018; Wu et al. 2023; Wang et al. 2023) often fall into one of two problematic extremes:

Specifically, some recommendation methods tend to be overly aggressive, overestimating action values for unseen treatments (Lee et al. 2024; Wu et al. 2023; Niraula et al. 2021), such as recommending excessive medication doses that may harm patients. To solve this, approaches like Conservative Q-Learning (Ghasemi et al. 2025; Lee et al. 2024) and filtering rare actions (Drudi et al. 2024) have been employed. However, such penalization-based methods may result in the opposite problem: leading to overly conservative policies, which prevent rare but potentially effective interventions.

To resolve this dilemma, one promising direction is integrating domain knowledge, such as explicit symbolic rules derived from expert behaviors or prior knowledge (Yang et al. 2018; Lu et al. 2018; Zhang and Sridharan 2022; Hoang et al. 2024). Such domain knowledge provides an effective middle ground: rule-guided exploration considers established safety boundaries while still allowing explorative decision-making within these boundaries. These "guardrails" prevent dangerous actions yet enable discovery of novel, effective treatments that penalization-based methods may overlook. To our knowledge, no prior work has introduced such a rule-based RL framework for clinical treatment tasks, positioning our approach as a pioneering contribution to decision-making systems in safety-critical medical applications.

Why This Action? A trustworthy clinical AI system requires a clear reasoning process that shows entire decision pathway, from initial patient states to final treatment recommendations. Such reasoning process can help physicians fully comprehend how and why specific treatment is proposed for this patient. However, most models mainly optimize outcome-driven objectives (e.g., 90-day mortality) (Yin et al. 2022; Komorowski et al. 2018; Fatemi et al. 2021; Ghasemi et al. 2025; Wu et al. 2023), focusing on statistical performance rather than mechanistic understanding of disease. Even some (Ghasemi et al. 2025; Lee et al. 2024) are supported by post-hoc explanation methods like feature importance, providing only surface-level information, not clinically meaningful interpretations (Hein, Udluft, and Runkler 2018; Lage et al. 2019; Ghasemi et al. 2025). In contrast, our design allows tracing the reasoning pathway leading to a treatment recommendation.

Understanding why an action is recommended is critical, *as metaphorically echoed by the ancient Oracle's wisdom,* ***"The power of Delphi was never in the answer — it was in the why."*** To achieve this, we employ neuro-symbolic approaches, which integrate neural networks' powerful data processing capabilities with the interpretable reasoning structures of symbolic systems (Zhang and Sheng 2024; Verhagen et al. 2022; Yang et al. 2025). Our proposed neuro-symbolic RL method, Delphi, can structurally answer

the three questions and offer the following contributions:

- **Causality-Aware State Modeling**: Delphi constructs personalized causal graphs dynamically adapted to each patient. This explicit causal structure enables mechanism-aware reasoning that reflects inter-subsystem pathological interactions. This provides an individualized understanding of each patient, directly addressing *"Why for this patient"*.
- **Adaptive Neuro-Symbolic Safety Constraints**: We use both clinical and data-driven rules as trainable safety constraints during learning. These rules serve as patient-specific safety checks, directly answering *"Why is it safe?"*.
- **Neuro-Symbolic Integration with LLM-Driven Explanations**: Delphi combines neural Q-values with rules to make decisions. This fusion is dynamic to keep learning flexible while adding clear, rule-based logic. A large language model explains each action by tracing its reasoning steps, offering transparent and clinically grounded answers to *"Why this action?"*.
- We have shown that Delphi outperforms over existing RL baselines, and we have conducted the *first pilot evaluation involving 8 ICU physicians* across varying levels of seniority. Blinded assessments along six clinical dimensions indicated Delphi's improvements in aspects such as adoption rate, effectiveness, satisfaction.

Methodology

Delphi introduces a novel neuro-symbolic causal RL framework for interpretable, safe, and individualized treatment recommendations (Fig. 1). Specifically, it consists of: 1) causality-aware state modeling with patient-specific causal graphs (Section 1); 2) adaptive, differentiable symbolic rule constraints for safety (Section 2); 3) final actions selected through neuro-symbolic fusion and explained by LLM (Section 3).

Why for This Patient at This Time? Interpretable Causality-Aware State Modeling

Clinically Guided Discretization and Subsystem To reason effectively about personalized treatment, an RL agent must begin with a clear and structured understanding of the patient's current condition. This needs to go beyond black-box vectors and instead captures the physiological relationships among clinical variables in a way that is both clinically meaningful and interpretable.

We represent raw clinical observations as $o_t \in \mathbb{R}^n$, and discretize them into a categorical state vector $D(o_t) = [d_{t,1}, \ldots, d_{t,n}]$, where each $d_{t,j} \in \{1, \ldots, K_j\}$ denotes a clinically meaningful category (e.g., "low", "normal", "high" blood pressure), based on guideline-derived thresholds. These discrete variables are organized into predefined medical subsystems. This design is inspired by Sepsis-3 (Rhodes et al. 2017) and SIRS (Marik and Taeb 2017) criteria, which emphasize as a disease affecting multiple organs. By grouping variables according to organ system, we enable structured, subsystem-specific reasoning that mirrors clinical practice—where **physicians interpret laboratory values and vital signs in relation to their specific organ systems rather than as standalone measurements**. This subsystem architecture provides a medically grounded foundation for downstream modules, including GNN-based state embedding (Section 1.3) and LLM-generated explanations (Section 3.2), thereby enhancing both interpretability and alignment with physician reasoning.

Dynamic Causal Graphs Based on Sepsis Subgroups To model patient-specific pathophysiology, we establish a causal graph framework grounded in sepsis subgroups. We construct sepsis subgroups based on organ-specific dysfunction criteria from the Sepsis-3 consensus [31], which defines sepsis by dysregulated immune response leading to life-threatening organ dysfunction. Each subgroup corresponds to a major organ system (e.g., coagulation) and captures the pathological relationships relevant to its associated dysfunction. Following this principle, we define subgroups based on the six major systems' dysfunction that Sepsis-3 focuses on: *neurological, respiratory, coagulation, hepatic, cardiovascular*, and *renal systems*. For each subgroup, we learn a specific causal graph structure $E^{(c)} \in \{0,1\}^{N \times N}$ using MIMIC-III samples that meet the SOFA criteria standard (e.g., neurological dysfunction is defined by Glasgow Coma Scale $\leq$ 14). These graphs are established offline via Peter–Clark Algorithm (Spirtes and Glymour 1991).

Each causal graph of a subgroup exhibits a distinct causal structure, reflecting the differences in underlying pathophysiological mechanisms.

At each time step t, each patient is assigned to a sepsis subgroup c_t, which selects the corresponding pre-learned subgroup graph $E^{(c_t)}$. To personalize this graph, we apply edge modulation: $E_t[i,j] = E^{(c_t)}[i,j] \cdot \alpha_{ij}(s_t)$, where $\alpha_{ij}(s_t) \in [0,1]$ is learnable and represents the patient-specific influence under the current patient state s_t. In a complete causal graph with N clinical variables, if we compute a unique modulation factor $\alpha_{ij}(s_t)$ for each edge, the total number of parameters grows quadratically, resulting in high computational cost. To address this, we introduce a subsystem-based edge grouping mechanism. We group edges by medical subsystem pairs $G_{k,l} = \{(i,j) \mid i \in \text{sys}_k, j \in \text{sys}_l\}$, where i belongs to subsystem k and j belongs to subsystem l. For each subsystem pair, we then learn a system-shared modulation factor $\alpha_{G_{k,l}}(s_t) = \sigma(f_{k,l}(h(s_t)))$. This means that for any edge (i,j) where $i \in \text{sys}_k$ and $j \in \text{sys}_l$, we have $\alpha_{ij}(s_t) = \alpha_{G_{k,l}}(s_t)$; therefore, all edges between any two given medical subsystems are collectively modulated by this $\alpha_{G_{k,l}}(s_t)$ factor. To ensure training stability, a progressive graph updating strategy is employed: $E_t = \beta \cdot E_t + (1-\beta) \cdot E_{t-1}$, where E_t is the personalized graph obtained from edge modulation, and $\beta \in (0,1)$ is a learnable rate controller.

State Representation by GNN We combine three components to form the patient's state representation: discretized clinical observations $D(o_t)$, a patient-specific causal graph E_t, and the assigned subgroup identifier c_t. We process this state using a Graph Neural Network (GNN) with a mes-

sage passing mechanism designed to model causal chains. At each GNN layer l, node i (representing variable $d_{t,i}$ from the discretized clinical observations $D(o_t)$) updates its state by aggregating messages from its causal parents j in E_t: $h_{t,i}^{(l+1)} = \sigma\left(W_1^{(l)} h_{t,i}^{(l)} + W_2^{(l)} \sum_{j:E_t[j,i]=1} h_{t,j}^{(l)} + b^{(l)}\right) + \gamma^{(l)} h_{t,i}^{(0)}$. Here, $h_{t,i}^{(0)}$ is the initial feature representation of variable $d_{t,i}$, derived from its clinically meaningful discretized value in $D(o_t)$ (e.g., via an embedding layer).

Stacking L GNN layers enables each node to recursively propagate and receive messages along directed edges in the patient-specific causal graph E_t. After L layers, each node incorporates information from all its upstream ancestors up to L hops away. This mechanism allows the embedding of a physiological variable, for example, "low creatinine" reflects not only its current value but also the influence from upstream causes. The message passing over the causal path would be like *hypovolemia → hypotension → renal hypoperfusion → elevated creatinine*. This helps Delphi answer "Why for this patient?" with physiologically-grounded reasoning that mirrors clinical diagnostic thinking.

Why Is It Safe? Adaptive Rule Constraint Learning

Trainable Symbolic Rules To ensure Delphi's decisions are both safe and effective, we incorporate two types of symbolic rules: guideline-based rules ($R_{\text{guideline}}$) and behavioral rules ($R_{\text{behavioral}}$). This hybrid structure includes both clinical expertise and data-driven insights. $R_{\text{guideline}}$ that are directly derived from authoritative medical consensus defines safety boundaries, which help prevent extreme or unsafe actions. $R_{\text{behavioral}}$ extracted from a pretrained Q-network captures high-reward treatment patterns learned from large-scale historical data. By analyzing the Q-network's decision logic, we distill these effective strategies into interpretable Horn rules (e.g., *IF state = "hypotensive" THEN recommend medium vasopressor dosage*), enabling soft policy shaping that guides the model toward effective treatments. By serving as explicit safety guardrails, these rules help Delphi's clinical decisions stay aligned with safe practice.

Specifically, $R_{\text{guideline}}$ is directly translated into Horn rules from relevant clinical guidelines provided by physicians (see details in Appendix §2). For $R_{\text{behavioral}}$, we analyze Deep Q Network (DQN) pretraining trajectories to extract candidate Horn rules. The rule evaluation process has two steps: 1) **Frequency-based filtering**: We assign each rule a confidence score based on its occurrence frequency in the data. Rules appearing rarely (below a threshold of 0.03) are discarded; 2) **Gradient-guided feature selection**: We compute Q-value gradients $\nabla_{x_i} Q(s_t, a)$ from the pretrained DQN to identify which physiological features most strongly influence each action decision. Features with larger gradient magnitudes are prioritized when constructing rules, ensuring the rules capture key clinical decision factors.

Once these rules are ready, we move onto evaluating how well they match a patient's current condition and how strongly they should influence the current decision-making process. For example, a Horn rule may be: **IF** Mean Arterial Pressure(MAP) = low **AND** Lactate = high **THEN** recommend medium vasopressor. Here, the "IF" part defines two premises: $P_{k,1}$ is "MAP = low" and $P_{k,2}$ is "Lactate = high". Taking a patient with MAP = 55 mmHg (very low) and Lactate = 3.8 mmol/L (moderately high) as an example, Delphi calculates a score indicating how closely the patient's condition matches the condition's requirement. For instance, "MAP = low" might be defined with a sigmoid function centered at 65 mmHg, yielding $p_{k,1}(s_t) = \sigma((65-55)/5) = 0.9$ for our patient with MAP = 55 mmHg. Similarly, "Lactate = high" might use a sigmoid centered at 2.5 mmol/L, giving $p_{k,2}(s_t) = \sigma((3.8-2.5)/2) = 0.7$ for the Lactate value of 3.8 mmol/L. These individual scores are aggregated via a product T-norm to compute how well the patient's current state jointly satisfies all conditions in the rule's IF part: $\prod_{j=1}^{n_k} p_{k,j}(s_t)$. In our example, this overall score is $0.9 \times 0.7 = 0.63$.

For each rule, we define a symbolic rule support $S_k(s_t)$, which plays a crucial role in our neuro-symbolic decision-making process (Section 3.1). It serves as the direct rule-driven input that influences the final action selection. Symbolic rule support is calculated as $S_k(s_t) = \omega_k \cdot \left(\prod_{j=1}^{n_k} p_{k,j}(s_t)\right)$, where $\omega_k \in [0,1]$ is the learnable confidence weight.

For example, if one rule has a confidence weight of $\omega_k = 0.8$ (indicating it appears frequently in successful treatment trajectories), the symbolic rule support would be $S_k(s_t) = 0.8 \times 0.63 = 0.504$. This score quantifies how strongly the rule recommends the action for this specific patient.

Why This Action? Neuro-Symbolic Decision Making and Explanation

Joint Neuro-Symbolic Decision Making The final treatment decision, represented by the recommended action a, is determined by a confidence-weighted gating mechanism that integrates outputs from both the neural DQN and the symbolic rule system, dynamically balancing data-driven optimization and rule-based constraints (see details in Appendix §3).

Explanation Generation For each recommended action, the system generates a structured and informative explanation (see Fig. 2) using an LLM (DeepSeek-V3), including: 1) *Patient Context:* The state $s_t = (V_t, E_t, c_t)$ encodes clinical measurements V_t, the dynamically modulated causal graph E_t, and the patient's clinical subgroup c_t. 2) *Decision Evidence:* Neural Q-values $Q_{\text{DQN}}(s_t, a)$ and the set of activated symbolic rules with the learned rule confidence. 3) *Clinical Alignment:* the rules from relevant guidelines supporting the recommendation, ensuring alignment with medical standards. Each decision is traceable—explaining why this action is chosen, for this patient, at this time.

Experiments

Datasets and Baselines We use de-identified ICU records from MIMIC-III (Johnson et al. 2016), following the preprocessing protocol of (Komorowski et al. 2018). The dataset is split into 16,753 training and 4,189 test trajectories. For evaluation, we select three off-policy RL baselines: *AI*

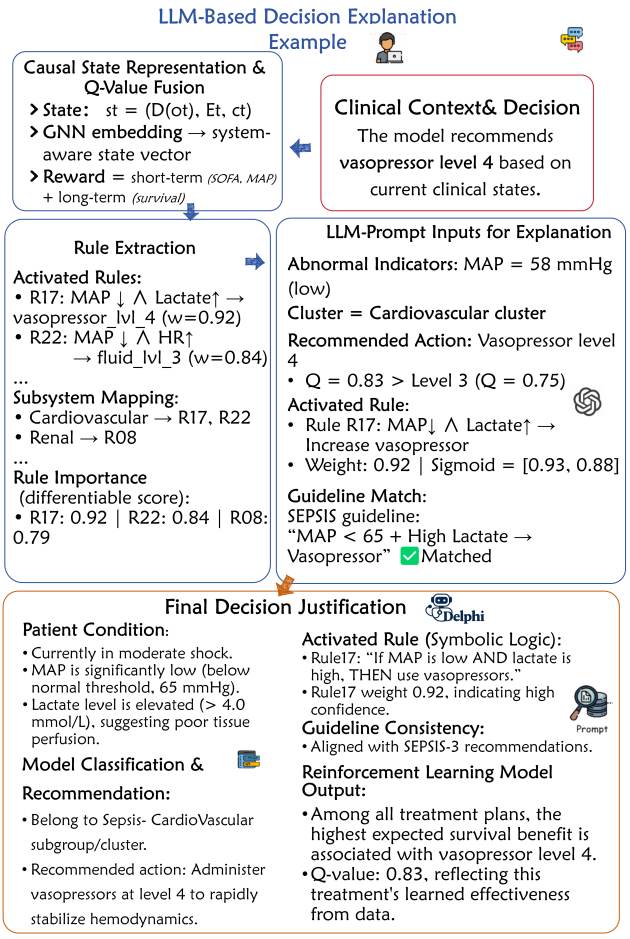

Figure 2: LLM-based decision explanation. The figure illustrates how the language model generates structured, patient-specific rationales aligned with the causal state and policy outputs.

Clinician (Komorowski et al. 2018), that demonstrates the feasibility of reinforcement learning in sepsis treatment; *WD3QNE* (Wu et al. 2023) that incorporates human expertise directly into the value function estimation to ensure clinical safety particularly with focus on less severe patient states. *DeD* (Fatemi et al. 2021) that enhances safety by active learning to avoid high-risk treatments. To ensure a fair comparison among all competing methods, we group patient states into multiple clusters (i.e., 750 and 100) via k-means++. Such discretization defines the finite state space required for reinforcement learning and provides a basis for comparing AI-generated and physician's treatment strategies (Komorowski et al. 2018).

Evaluations We adopt a comprehensive set of evaluation metrics following Luo et al. (2024), which employs more meaningful metrics than traditional methods (Komorowski et al. 2018; Wu et al. 2023), listed as follows. **1) Off-Policy Evaluation (OPE):** We report Weighted Importance Sampling (WIS) (Tokdar and Kass 2010), its truncated variant (WISt, with weight cap), Bootstrapped WIS (WISb, averaged over B=100 resamples), and Bootstrapped Truncated WIS (WISbt) that enhances stability by combining both. Doubly Robust (DR) (Jiang and Li 2016) is also included which combines importance sampling and value estimation, and is reliable as long as one of these two components is correct. **2) Treatment Discrepancy:** To assess the alignment between model-recommended treatments and those prescribed by clinicians, we use RMSE-IV (Wang et al. 2018) and RMSE-VASO (Huang, Cao, and Rahmani 2022), measuring the root mean squared error for two treatments, *i.e.*, IV fluid and vasopressor predictions. **3) Clinical Action Consistency:** We compute F1 scores at the patient level (P.F1) (Powers 2020; Luo et al. 2024) and time-step level (S.F1) (Luo et al. 2024) to evaluate categorical agreement with physician decision.

Competing Methods As shown in Table 1, Delphi generally outperformed AI Clinician (Komorowski et al. 2018) across different evaluation metrics, including Off-Policy Evaluation (OPE), Treatment Discrepancy, and Clinical Action Consistency. Even AI Clinician' WIS (0.9440) looks good, it failed on its evaluation variants (i.e., WISt, WISb, WISbt, and DR), showing negative or near-zero scores (e.g., WISt: –0.8436; DR: 0.1225). In contrast, Delphi consistently achieved positive results across all OPE metrics, clearly demonstrating its policies are reliably effective. Beyond OPE, Delphi also demonstrated superior performance in Treatment Discrepancy (e.g., lower RMSE-VASO of 0.2071 vs. AI Clinician's 0.2111) and Clinical Action Consistency (e.g., higher S.F1 score of 0.5601 vs. AI Clinician's 0.4719), indicating better alignment with physician practices.

This is because traditional approaches like AI Clinician initially employed 750 clusters for patient states which lacked clear clinical meaning, limiting models' ability to learn complex pathophysiological patterns. In contrast, Delphi abandons this hyperparameterized, arbitrary clustering method and introduces Causality-Aware State Modeling. It organizes clinically meaningful, discretized variables and dynamically builds patient-specific causal graphs, yielding a state representation that is both clinically grounded and interpretable. This allows Delphi to answer "Why for this patient?" and deliver more personalized and accurate decisions.

Compared to the WD3QNE model, Delphi also outperformed in OPE metrics for the same reasons discussed earlier—better state modeling. Our model also achieved higher scores in clinical action consistency metrics (P.F1 and S.F1), primarily due to our introduction of Adaptive Rule Constraints. These constraints integrate established medical protocols directly into the policy learning loop, safeguarding Delphi's decision-making process by preventing unsafe or overly aggressive interventions.

DeD avoids high-risk treatments by selecting actions with the lowest estimated danger, rather than optimizing for expected rewards (Fatemi et al. 2021). DeD achieves moderate OPE scores. However, since both DeD and Delphi are safety-prioritized, and Delphi works well because it uses

Models	(1) Off-Policy Evaluation (OPE) ↑					(2) Treatment Discrepancy ↓		(3) Clinical Action Consistency ↑	
	WIS	WISt	WISb	WISbt	DR	RMSE_{IV}	RMSE_{Vaso}	PF1	SF1
Delphi (Ours)	**1.4700**	**1.2800**	0.7824	0.7147	**1.2900**	0.7862	**0.2071**	**0.5497**	**0.5601**
AI Clinician (750 clusters)	0.9440	-0.8436	-0.3463	-0.2218	0.1225	**0.6749**	0.2111	0.5098	0.4719
AI Clinician (100 clusters)	0.9141	0.8233	**0.8663**	0.7790	-0.0408	0.7299	0.2261	0.4908	0.4571
WD3QNE (750 clusters)	0.8120	0.7264	0.7881	0.6962	0.8440	0.8830	0.2301	0.3347	0.3037
WD3QNE (100 clusters)	0.8860	0.0158	0.6859	0.7292	0.6851	0.9000	0.2308	0.3325	0.3023
DeD (750 clusters)	0.7841	0.7809	0.7877	**0.7898**	0.6946	0.8329	0.2181	0.3661	0.3272
DeD (100 clusters)	0.7633	0.7531	0.7648	0.7652	0.7210	0.8782	0.2257	0.3610	0.3197

Table 1: Comparison of Delphi with existing baselines.

Models	(1) Off-Policy Evaluation (OPE) ↑					(2) Treatment Discrepancy ↓		(3) Clinical Action Consistency ↑	
	WIS	WISt	WISb	WISbt	DR	RMSE_{IV}	RMSE_{Vaso}	PF1	SF1
w/o Causal Graph	0.9844	0.5877	0.6093	0.6837	0.3700	**0.4692**	0.5892	**0.6290**	0.3844
w/o Rule Reasoning	0.9066	0.8573	**0.7983**	0.6223	0.7100	0.8829	0.4032	0.5500	0.3999
w/o GNN	0.7889	0.6983	0.6945	0.5882	0.1700	0.7696	0.7631	0.4434	**0.6478**
Delphi (Full)	**1.4700**	**1.2800**	0.7824	**0.7147**	**1.2900**	0.7862	**0.2071**	0.5497	0.5601

Table 2: Ablation study of Delphi.

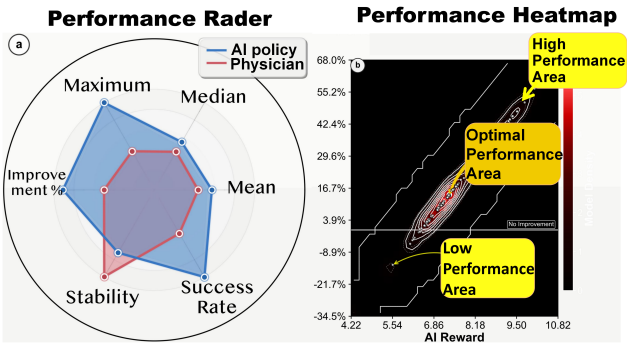

Figure 3: Performance comparison between Delphi and physician policies. (a) Radar chart summarizes six evaluation metrics. (b) Density heatmap shows trained model distribution by AI reward and improvement over physicians.

clinically meaningful state modeling. Thus we wanted to see if DeD would perform better after modifying its state representation. Indeed, when applied with a 6-cluster configuration (aligned with Sepsis-3 defined organ dysfunction subgroups, as in Delphi), DeD achieved a strikingly high Doubly Robust (DR) value (3.0245) in our reproduction, where we strictly followed its risk-tagged Q-values to choose actions. This unexpectedly strong DR score reveals a deeper mechanism: selecting actions based on the worst-case avoidance is naturally effective. DeD's cautious action choices reduce the divergence between its learned policy and the data's behavior policy, resulting in higher DR scores. This finding directly supports DeD's core hypothesis: in high-stakes clinical settings, safety-driven methods can be as effective as reward-driven strategies (Komorowski et al. 2018; Fatemi et al. 2021).

Ablation Studies As shown in Table 2, when removing either the causal graph structure ("w/o Causal Graph") or GNN ("w/o GNN"), Delphi's OPE performance dropped largely (e.g., WIS dropped from 1.4700 to 0.9844 and 0.7889, respectively; DR dropped from 1.2900 to 0.3700 and 0.1700, respectively). This demonstrates that both the causal graph structure and the GNN are essential for modeling patient pathophysiology to achieve optimal performances. Furthermore, when Adaptive Rule Constraints were removed, the model's performance also deteriorated (e.g., WIS dropped from 1.4700 to 0.9066). This highlights the importance of integrating adaptive rules into the learning process. These rules serve as patient-specific safety checks, not only ensuring the clinical safety but also improving consistency with physician practices, thereby effectively answering the crucial question of "Why is it safe?".

Delphi is further evaluated in Fig. 3a through a radar chart comparing its policy with physicians' policy across six normalized dimensions (scaled 0–1): Mean Reward, Median Reward, Maximum Reward, Success Rate (AI outperforms physicians), Stability (1 - normalized standard deviation of reward), and Improvement Percentage (normalized gain over physician). Fig. 3b shows the relationship between AI reward and improvement over the physician policy, with most good models grouped in a high-density area (reward: 8.0–9.5, improvement: 20–45%).

Interpretability and Safety We introduce a blinded assessment involving eight ICU physicians of varying seniority. Each physician independently rated treatment decisions made by Delphi and historical physicians for 50 real-world sepsis patients at each time step. Decisions were evaluated along six clinical dimensions (i.e., Adoption, Effectiveness, Safety, Satisfaction, Understanding, Trust) using a 7-point Likert scale (see Fig. 4).

Explanation Clarity — Why taking this action? Delphi scored **5.40** in understandability, outperforming historical

physician decisions (**4.96**, +8.9%). This reflects the strength of Delphi's LLM-based explanations and neural-symbolic Q fusion. The former offers clear justifications, while the latter provides evidence of clinically-aligned reasoning.

Trust and Adoption — Why for this patient? Delphi achieved a **5.20** trust score (vs. **4.78**, +8.8%) and a higher adoption rate (**83.75%** vs. **78.00%**). These improvements are driven by its causal-aware state modeling, which enables patient-specific reasoning. Delphi's recommendations were more frequently accepted in complex cases (5, 6, 10), demonstrating its ability to translate clinical data into actionable insights, especially when physicians face challenging scenarios.

Safety — Why is this action safe? Delphi achieved a safety score of 5.01, marking a strong improvement over historical physician decisions (4.54). Physicians have evaluated safety based on indicators such as guideline violations, potential shifts in mortality risk, and adherence to dose limits. For instance, in Case 7, a stress test scenario involving hypotension and renal failure, Delphi avoided unsafe interventions and aligned with critical medical thresholds, thereby reducing potential clinical errors.

Clinical Impact — Delphi also scored higher in both **effectiveness (4.90)** and **satisfaction (5.03)**. Crucially, physicians spent more time reviewing Delphi's suggestions (**12.1s** vs. **4.18s**). This is a positive sign: Delphi's structured medical insights, grounded in pathological mechanisms and causal reasoning, actively prompt physician reflection. Clinicians integrate Delphi's suggestions into their own decision-making, exemplifying genuine co-diagnostic reasoning between AI and physicians. Coupled with Delphi's higher adoption rates, trust scores, and satisfaction levels, this deeper form of interaction suggests that Delphi fosters a human-machine collaborative clinical environment, supporting more transparent and robust decision-making pathways.

Qualitative Feedback and Evaluation — We conducted a detailed follow-up questionnaire on why physicians did not adopt Delphi's recommended treatment plans. The main reasons included: perceived oversimplicity conflicting with clinical judgment (**25%**), insufficient clarity of explanation (**20%**), omission of key factors (**30%**), personal experience favoring alternatives (**15%**), and perceived safety concerns (**10%**). These responses suggest that despite Delphi's strong safety performance, improvements are needed in explanation clarity and treatment design, particularly given the current reliance on discrete action options. Additionally, a 30-item blinded evaluation compared physician perceptions of treatment recommendations from Delphi (with explanations) and Vanilla RL (which offers no explanations (Komorowski et al. 2018)) under identical patient scenarios. Delphi consistently received higher ratings across all perceptual dimensions: effectiveness (**4.5** vs. **4.2**), safety (**4.0** vs. **3.8**), satisfaction (**4.2** vs. **4.0**), understandability (**4.8** vs. **4.5**), and trust (**4.6** vs. **4.3**). These findings highlight Delphi's advantages in perceived reliability, safety, and overall physician willingness to adopt its recommendations over a non-explaining RL baseline.

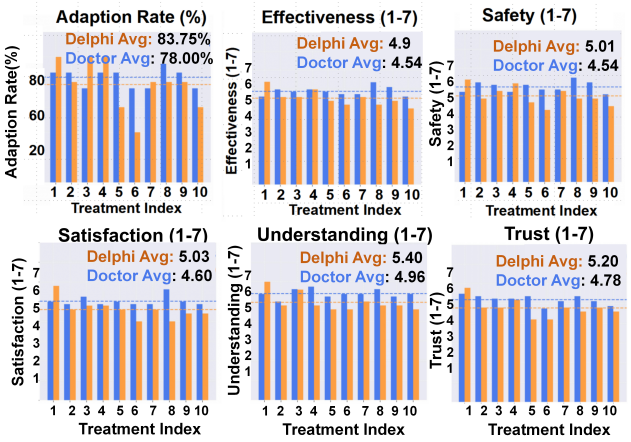

Figure 4: Blind Evaluation of Delphi vs. historical physician decisions across 10 treatments on 6 dimensions.

Conclusion

Delphi represents a big step toward making clinical reinforcement learning individualized, safe, and interpretable, systematically answering the three "why" questions. Delphi outperforms existing methods across both quantitative evaluation metrics and qualitative physician assessments. Notably, this is the first blinded evaluation of an interpretable RL system in healthcare, where physicians of varying expertise levels rated Delphi highly across multiple dimensions.

This outperformance stems from Delphi's innovative design, which addresses the key limitations of prior approaches relying on clinically meaningless state representations. Delphi uses clinically-guided causal state representation, developed in discussions with sepsis experts and aligned with how physicians assess organ dysfunction. Equally important, Delphi ensures safety through adaptive symbolic rule constraints. Together, these two designs not only improve accuracy but also provide patient-specific justification of treatments .

To conclude, by delivering structured, traceable reasoning for each recommendation, Delphi enables clinicians to trace, verify, and challenge AI-driven decisions when needed. This establishes a foundation of accountability, transforming AI from a black-box predictor into a transparent and collaborative clinical assistant.

References

Drudi, C.; Mollura, M.; Li-wei, H. L.; and Barbieri, R. 2024. A reinforcement learning model for optimal treatment strategies in intensive care: assessment of the role of cardiorespiratory features. *IEEE Open Journal of Engineering in Medicine and Biology*, 5: 806–815.

Fatemi, M.; Killian, T. W.; Subramanian, J.; and Ghassemi, M. 2021. Medical dead-ends and learning to identify high-risk states and treatments. *Advances in Neural Information Processing Systems*, 34: 4856–4870.

Fleischmann-Struzek, C.; and Rudd, K. 2023. Challenges

of assessing the burden of sepsis. *Medizinische Klinik-Intensivmedizin und Notfallmedizin*, 118(Suppl 2): 68–74.

Ghasemi, P.; Greenberg, M.; Southern, D. A.; Li, B.; White, J. A.; and Lee, J. 2025. Personalized decision making for coronary artery disease treatment using offline reinforcement learning. *npj Digital Medicine*, 8(1): 99.

Glanois, C.; Weng, P.; Zimmer, M.; Li, D.; Yang, T.; Hao, J.; and Liu, W. 2024. A survey on interpretable reinforcement learning. *Machine Learning*, 113(8): 5847–5890.

Hein, D.; Udluft, S.; and Runkler, T. A. 2018. Interpretable policies for reinforcement learning by genetic programming. *Engineering Applications of Artificial Intelligence*, 76: 158–169.

Hoang, T. L.; Sbodio, M. L.; Galindo, M. M.; Zayats, M.; Fernandez-Diaz, R.; Valls, V.; Picco, G.; Berrospi, C.; and Lopez, V. 2024. Knowledge enhanced representation learning for drug discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 10544–10552.

Huang, Y.; Cao, R.; and Rahmani, A. 2022. Reinforcement learning for sepsis treatment: A continuous action space solution. In *Machine Learning for Healthcare Conference*, 631–647. PMLR.

Jiang, N.; and Li, L. 2016. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, 652–661. PMLR.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1): 1–9.

Komorowski, M.; Celi, L. A.; Badawi, O.; Gordon, A. C.; and Faisal, A. A. 2018. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11): 1716–1720.

Lage, I.; Lifschitz, D.; Doshi-Velez, F.; and Amir, O. 2019. Exploring computational user models for agent policy summarization. In *IJCAI: proceedings of the conference*, volume 28, 1401.

Lee, H. Y.; Chung, S.; Hyeon, D.; Yang, H.-L.; Lee, H.-C.; Ryu, H. G.; and Lee, H. 2024. Reinforcement learning model for optimizing dexmedetomidine dosing to prevent delirium in critically ill patients. *npj Digital Medicine*, 7(1): 325.

Lu, K.; Zhang, S.; Stone, P.; and Chen, X. 2018. Robot representation and reasoning with knowledge from reinforcement learning. *arXiv preprint arXiv:1809.11074*.

Luo, Z.; Pan, Y.; Watkinson, P.; and Zhu, T. 2024. Reinforcement Learning in Dynamic Treatment Regimes Needs Critical Reexamination. arXiv:2405.18556.

Marik, P. E.; and Taeb, A. M. 2017. SIRS, qSOFA and new sepsis definition. *Journal of thoracic disease*, 9(4): 943.

Niraula, D.; Jamaluddin, J.; Matuszak, M. M.; Haken, R. K. T.; and Naqa, I. E. 2021. Quantum deep reinforcement learning for clinical decision support in oncology: application to adaptive radiotherapy. *Scientific reports*, 11(1): 23545.

Powers, D. M. 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Rhodes, A.; Evans, L. E.; Alhazzani, W.; Levy, M. M.; Antonelli, M.; Ferrer, R.; Kumar, A.; Sevransky, J. E.; Sprung, C. L.; Nunnally, M. E.; et al. 2017. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock: 2016. *Intensive care medicine*, 43(3): 304–377.

Rudd, K. E.; Johnson, S. C.; Agesa, K. M.; Shackelford, K. A.; Tsoi, D.; Kievlan, D. R.; Colombara, D. V.; Ikuta, K. S.; Kissoon, N.; Finfer, S.; et al. 2020. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *The Lancet*, 395(10219): 200–211.

Singer, M.; Deutschman, C. S.; Seymour, C. W.; Shankar-Hari, M.; Annane, D.; Bauer, M.; Bellomo, R.; Bernard, G. R.; Chiche, J.-D.; Coopersmith, C. M.; et al. 2016. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *Jama*, 315(8): 801–810.

Spirtes, P.; and Glymour, C. 1991. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1): 62–72.

Tokdar, S. T.; and Kass, R. E. 2010. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1): 54–60.

Verhagen, R. S.; Mehrotra, S.; Neerincx, M. A.; Jonker, C. M.; and Tielman, M. L. 2022. Exploring Effectiveness of Explanations for Appropriate Trust: Lessons from Cognitive Psychology. *arXiv preprint arXiv:2210.03737*.

Wang, G.; Liu, X.; Ying, Z.; Yang, G.; Chen, Z.; Liu, Z.; Zhang, M.; Yan, H.; Lu, Y.; Gao, Y.; et al. 2023. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29(10): 2633–2642.

Wang, L.; Zhang, W.; He, X.; and Zha, H. 2018. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2447–2456.

Wu, X.; Li, R.; He, Z.; Yu, T.; and Cheng, C. 2023. A value-based deep reinforcement learning model with human expertise in optimal treatment of sepsis. *NPJ Digital Medicine*, 6(1): 15.

Yang, F.; Lyu, D.; Liu, B.; and Gustafson, S. 2018. Peorl: Integrating symbolic planning and hierarchical reinforcement learning for robust decision-making. *arXiv preprint arXiv:1804.07779*.

Yang, X.; Shao, J.; Guo, L.; Zhang, B.; Zhou, Z.; Jia, L.; Dai, W.; and Li, Y. 2025. Neuro-Symbolic Artificial Intelligence: Towards Improving the Reasoning Abilities of Large Language Models. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*, 10770–10778.

Yin, C.; Liu, R.; Caterino, J.; and Zhang, P. 2022. Deconfounding actor-critic network with policy adaptation for dynamic treatment regimes. In *Proceedings of the 28th ACM*

SIGKDD Conference on Knowledge Discovery and Data Mining, 2316–2326.

Zhang, S.; and Sridharan, M. 2022. A survey of knowledge-based sequential decision-making under uncertainty. *AI Magazine*, 43(2): 249–266.

Zhang, X.; and Sheng, V. S. 2024. Neuro-symbolic AI: Explainability, challenges, and future trends. *arXiv preprint arXiv:2411.04383*.