Data-Driven Structural Optimization of Descriptive Norms: Bridging Micro and Macro Dynamics in Medical Contexts

Chao Li[0009−0000−1233−3281], Ilia Derevitskii[0000−0002−1690−9812], and Sergey Kovalchuk[0000−0001−8828−4615]

ITMO University, Saint Petersburg, Russia
316325@niuitmo.ru, ilyaderevitskiy@gmail.com, kovalchuk@itmo.ru

Abstract. This paper presents a data-driven framework for optimizing descriptive norms in medical contexts, bridging micro-level agent behaviors and macro-level dynamics using partial differential equations (PDEs) and reinforcement learning. Descriptive norms are modeled as opinion distributions in a continuous space, evolving through diffusion and migration terms influenced by perceptual kernels and external potentials. The approach extends prior models by integrating Gaussian Mixture Models (GMMs) from real-world COVID-19 data. Top-down optimization employs Q-learning to align opinion distributions with target GMMs, reshaping clinical propensities (e.g., azithromycin usage) post-guidelines, achieving 82% similarity and faster convergence than random parameters. Bottom-up optimization uses causal effects from double machine learning on indicators like mortality, lymphocyte count, and consciousness decline, guiding PDE parameters to emergent norms that outperform baselines (e.g., 36.8% reduced consciousness decline).

Keywords: Norm optimization · Descriptive norm · PDEs · Complex systems · Autonomous multi-agents · Medical scenario

1 Background

In works [15, 5], based on the structure of agents' subjective perception of objective norms, formal definitions of descriptive norms and general models of their propagation and sharing were proposed. These works modeled the dynamics of descriptive norms based on hypothetical spatial topologies between agents, lacking more precise pattern representation of collective descriptive norms based on computable opinion spaces. Although human thinking has opacity, the dynamics of descriptive norms between agents can be directly computed at the data level within the opinion space.

Therefore, we use auto-aggregation mathematical models [11, 3, 8, 19, 9] to extend the model of descriptive norms, thereby providing direct computable methods for the structure and temporal optimization of such norms. These auto-aggregation mathematical models successfully characterize different aspects of collective dynamics, but the use of PDEs and transport equations within them,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_31

although strictly proven, has no direct connection to the bottom-up emergence or top-down macroscopic regulation of descriptive norms. Therefore, we extend the mathematical methods within them and combine them with the dataset-based formal model of descriptive norms in [15].

As the goal of this paper is the optimization of descriptive norms, we optimize the definition part related to Subjective Individual Norm Perception in [14] that has been empirically validated on datasets but is not suitable for collective norm reasoning in the opinion space.

2 Model

2.1 Mathematical model of descriptive norm

The PDEs part of the computable model of descriptive norms remains consistent with the previous work [14]. However, we no longer simply update parameters based on the Wasserstein distance between the Gaussian mixture distributions of SINP and OBJ. Instead, we systematically introduce agent-environment interactions in the opinion space and data-driven dynamic methods. Here, the descriptive norm reasoning of individual agents is reflected in the interaction between practice tendencies calculated by GMMs and practice outcome indicators (e.g., medical outcome indicators). Collective norm perception is embodied in the interaction kernel under the PDEs model.

$P(x,t)$ represents the population size holding opinion x at time t. Our auto-aggregation mathematical model [9, 11, 3, 19, 8] extends the function $P(x,t)$ describing opinion popularity dynamics to describe the group's violation and conformity to collective descriptive norms. Descriptive norms propagate and are shared within the collective, manifested as changes in opinion popularity.

The model assumes a constant total population, with distribution changes only due to diffusion and migration: diffusion reflects random opinion fluctuations, while migration represents directed active opinion shifts induced by social influence. Specifically, the mathematical model is based on the following assumptions: individuals only perceive information from neighboring regions in opinion space [2]; individuals not only learn norms but also decide when to comply, thereby adjusting perception range and direction [12, 13]; different individuals have subjective perceptions of "what constitutes normal behavior" [5]; the direction of group opinion movement forms a dynamic interaction between norm perception and group behavior.

We extend the classical transport equation framework [8, 6], and the mathematical description of P(x,t) is:

$$\frac{\partial P}{\partial t} = d\nabla^2 P - \nabla \cdot \left[P \left(cG(P) - \nabla V(x) \right) \right], \tag{1}$$

where $d\nabla^2 P$ is the diffusion term and $-\nabla \cdot [P\,(cG(P) - \nabla V(x))]$ is the migration term. $V(x) = k \cdot (x - x_{\text{target}})^2$ is the external potential field function, following the physical principle that force is the negative gradient of potential

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_31

energy, k is the potential strength parameter, and x_{target} is the target position. In the absence of an external force field with spatially dependent guidance (such as macroscopic norm regulation), the migration term automatically reduces to $-c\nabla\cdot(PG(P))$. $G(P)$ is the perceived gradient of popularity distribution, defined as:

$$G(P) = \int_{-\infty}^{\infty} P(x+y,t)g(y)dy, \tag{2}$$

$$g(y) = \frac{1}{2\mu}\cdot\frac{1}{\sqrt{2\pi}\sigma}\left(e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} - e^{-\frac{1}{2}\left(\frac{y+\mu}{\sigma}\right)^2}\right). \tag{3}$$

The external potential field is conditionally activated based on the number of opinion clusters present in the system, allowing for dynamic control of opinion aggregation behavior while maintaining the fundamental structure of the original transport equation model.

Within the PDE framework of generalized nonlocal gradients and interaction kernels, descriptive norms can be more generally defined in opinion space as the dynamic distribution of spatial collective opinions $p(x,t)$ (Equation 4). This encompasses both agents' internal conjectures or understandings of collective norms and the objective true norms themselves. The complete mathematical definition distinguishing subjective-objective special forms is provided in [14].

$$\mathcal{N}_{\text{desc}}(x,t) = \mathcal{F}[p(x,t), \Phi_{\text{sub/obj}}] \tag{4}$$

where $\mathcal{N}_{\text{desc}}(x,t)$ represents the descriptive norm at opinion position x and time t, $\mathcal{F}$ is a functional mapping that integrates the population opinion distribution $p(x,t)$ with the parameter space $\Phi_{\text{sub/obj}}$ that distinguishes between subjective perception and objective reality of norms within the generalized nonlocal gradient framework.

The generalized nonlocal gradient is a differential operator extended via kernel function. Equation 2 demonstrates that the perceptual gradient $G(P)$ represents the cross-correlation quantity between perception and kernel g – a classical interaction kernel approach in applied mathematics and physics [3, 6, 17, 19]. This perceptual kernel governs how individuals weight neighboring opinion popularity when evaluating gradients.

Where $\mu > 0$, the migration process simplifies to climbing behavior along the gradient of $P(x,t)$, manifesting as auto-aggregation that converges toward collective descriptive norms. When $\mu < 0$, the kernel polarity reverses ($g_\mu(y) = -g_{-\mu}(y)$), inducing auto-avoidance behavior that violates collective descriptive norms.

2.2 Descriptive norm structural optimization model

Model-free methods are suitable for our complex systems, such as descriptive norm dynamics processes described by partial differential equations. Here we

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_31

first describe the form of Q-learning that we use, and then describe how to use it for the optimization of descriptive norms.

Every MDP has at least one optimal policy and of the optimal policies for a given MDP, at least one is stationary and deterministic [16]. In the Markov Decision Process, we define the action-value function $Q(s, a)$ as: the total expected discounted reward obtained after executing action a from state s and then following the optimal policy. This function satisfies the Bellman optimality equation [16]:

$$Q^*(s, a) = \mathbb{E}\left[r + \gamma \max_{a'} Q^*(s', a') \mid s, a\right] \tag{5}$$

The state transition is implicitly defined by PDE simulation. Here, the model-free Q-learning directly learns the optimal Q function from environment interactions through Temporal Difference (TD) error:

$$Q_{k+1}(s, a) = Q_k(s, a) + \alpha\left[r + \gamma \max_{a'} Q_k(s', a') - Q_k(s, a)\right] \tag{6}$$

where the term in brackets is called the temporal difference error, which measures the difference between the current estimate and the newly observed target.

For terminal states, the update rule naturally simplifies to: $Q_{k+1}(s, a) = Q_k(s, a) + \alpha\left[r - Q_k(s, a)\right]$. The algorithm balances exploration and exploitation through the ϵ-greedy strategy, and improves data efficiency through "experience replay". Under appropriate conditions, Q-learning guarantees convergence to the optimal action-value function Q^*, thereby deriving the optimal policy $\pi^*(s) = \arg\max_a Q^*(s, a)$.

Top-down structural optimization The optimization of descriptive norms is still divided into two multi-scale parts - macro and micro (see Fig. 1). The macro part refers to top-down optimization, which means the issuance of guiding documents at the macro level leads to the reshaping of practice tendencies at the micro level. In our implementation, macro optimization transforms the difference between the GMM-represented target distribution $P_{\text{target}}(x)$ (representing the desired norm state) and the current collective opinion distribution $P(x, t)$ into reward signals through the Q-learning framework, guiding system parameter adjustment. Specifically, we define the reward function R to consider both distribution similarity and Wasserstein distance, enabling the system to learn optimal perceptual kernel parameters and potential field strength, thereby guiding the collective opinion distribution to the target norm state:

$$R = -\lambda_1 \cdot W(P(x,t), P_{\text{target}}(x)) + \lambda_2 \cdot \exp(-W(P(x,t), P_{\text{target}}(x))/\tau) - \lambda_3 \cdot k_{\text{p}} \tag{7}$$

where $W(\cdot, \cdot)$ denotes the Wasserstein distance between distributions, λ_1, λ_2, λ_3 are weighting coefficients that balance distribution alignment, similarity maximization, and control cost respectively, and k_{p} represents the strength of the

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_31

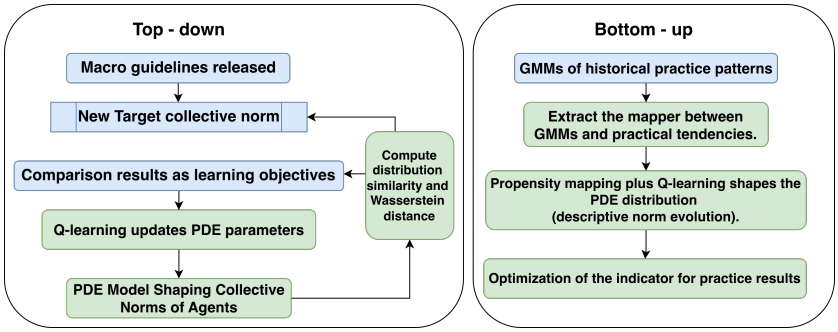

Fig. 1: Multi-scale descriptive norm structure optimization.

external potential field controlling opinion aggregation. This reward structure directly implements the Bellman optimality equation (Equation 5) by providing immediate feedback for each parameter selection step in the Markov Decision Process, enabling the Q-learning agent to discover optimal policy π^* that minimizes the distance between actual and target normative distributions.

Bottom-up norm structure optimization Bottom-up optimization requires the involvement of real-world practice data from the model's application scenarios. First, real-world practice data represented by Gaussian Mixture Models (GMMs) from historical periods is extracted to compute weighted average propensities for target practices, capturing empirical tendencies across operational clusters. These propensities inform a physical mapper that translates the evolved PDE opinion distribution P into a practice propensity score, which is then used to predict causal outcome metrics (e.g., core medical efficacy metrics, efficiency rates, performance levels) via a saturation-adjusted model grounded in observed effects. This propensity and predicted outcomes are integrated into the Q-learning reward function, where improvements over baseline metrics yield positive rewards, while deviations incur penalties, enabling the agent to iteratively shape the distribution through parameter selection (μ, σ, k) in PDE dynamics, ensuring data-driven convergence to optimized real-world indicators while preserving causal structures from empirical data.

The core transformations are captured in the following mappings:

$$\begin{aligned}\text{Propensity} = f_{\text{map}}(P, x) = m_{\min} + (m_{\max} - m_{\min}) \cdot \\ \sigma\left(\frac{\int P(x)x\,dx / \int P(x)\,dx - x_{\min}}{x_{\max} - x_{\min}}\right),\end{aligned} \tag{8}$$

where $\sigma(z) = 1/(1+e^{-10(z-0.5)})$ is a sigmoid for sensitivity, and $m_{\min}, m_{\max}$ are GMM-derived bounds.

$$r = \sum_i w_i \cdot g_i \left(h_i(\text{Propensity}) - b_i\right) + r_{\text{bonus}} - c_{\text{control}}, \tag{9}$$

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_31

where h_i predicts outcome metric i with saturation $1 - e^{-5 \cdot \text{Propensity}}$, g_i scales improvements/penalties relative to baseline b_i, w_i are weights, r_{bonus} rewards comprehensive gains, and c_{control} penalizes parameter changes.

This constitutes a bottom-up feedback loop: real data guides distribution evolution, experience replay enhances learning stability, and the process ultimately converges to practice metrics superior to the baseline while maintaining data-driven causal integrity.

3 Datasets and key medical indicators

We continue to use the real-world COVID-19 data from the Almazov National Medical Center as employed in [14]. Based on the three key events at the medical center—receiving the 7th edition guideline, first detection of the Alpha variant, and receiving the 10th edition guideline—we divided the two pandemic waves into 5 periods based on admission time, as established in [14]. The dataset exhibits two types of collective practice dynamics: (1) Following the release of national macro-level clinical guidelines, direct reshaping of micro-level practice tendencies occurs, which directly generates new collective descriptive norms; (2) Changes in medical reality caused by new variants prompt shifts in micro-level practice tendencies for clinical interventions, leading to emergent new descriptive norms within the collective that form a "violation" relationship with previous clinical guidelines.

All data processing procedures and computational methods for characterizing the two types of collective practice dynamics remain consistent with those in [14]. We continue to model the 33 control features across the five time periods as five 33-dimensional Gaussian Mixture Models (GMMs) through the continuous propensity field approach [10, 4, 1].

We use the significant change in azithromycin usage propensity after the release of v8/v9 guidelines as the target for our top-down experiment. Modeling of period 2's (from v7 guideline release to the end of Wave 1) propensity field shows hydroxychloroquine, azithromycin, and chloroquine were wrongly widespread in Wave 1 here, due to complex factors [20, 18, 7]. Post-Wave 1, v8/v9 guidelines flagged insufficient efficacy and cardiac risks for hydroxychloroquine/chloroquine, restricting their use. Thus, their propensity ranks fell sharply: hydroxychloroquine (1→28), chloroquine (8→29), azithromycin (4→27; co-administered with hydroxychloroquine).

We define Period 3 as the interval from the onset of the second wave of the pandemic to the first detection of the Alpha variant, and Period 4 as the interval from the Alpha variant to the emergence of the v10 guidelines. Our focus is on Periods 3 and 4, where emergent collective descriptive norms in medical control practices arise bottom-up due to the influence of the new variant on group tendencies.

No overlapping cases were found between Period 3 and Period 4. This indicates a cross-sectional difference-in-differences (DID) design, involving distinct patient cohorts in each period. We employed continuous propensity fields to

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_31

identify the most pronounced changes in control variables from Period 3 to 4, selecting prednisolone (26 to 10), methylprednisolone (19 to 13), and enoxaparin sodium (16 to 25) as key indicators.

Double machine learning (DML) was used to estimate the causal effects of these control variables on medical outcomes. The results passed false discovery rate (FDR) correction via the Benjamini-Hochberg procedure, accounting for potential confounding due to changes in glucocorticoid administration solely attributable to disease progression.

Among the causal effects of control columns on medical indicators, significant results (q-value < 0.01) were considered, including: level of consciousness decline; treatment outcome (where 1 indicates death and 0 indicates recovery); duration of observation and days since pandemic onset (both related to hospitalization duration); mean platelet volume (MPV); current disease duration; body temperature; absolute lymphocyte count; and lactate dehydrogenase (LDH) levels.

Among these, we selected changes in the following four indicators as key metrics for evaluating whether the descriptive norm structure of control columns is optimized. The critical reason for selecting these four indicators is that patient conditions vary significantly among individuals, and these indicators have relatively weak dependencies on other medical contexts, allowing their increases or decreases to directly assist in definitive clinical judgments: absolute lymphocyte count (increase/decrease), LDH levels (increase/decrease), level of consciousness decline (increase/decrease), and treatment outcome. An increase in absolute lymphocyte count represents a positive clinical signal, whereas a decrease signifies a negative risk indicator. A reduction in lactate dehydrogenase (LDH) levels indicates alleviation of tissue damage, while an elevation suggests intensified cellular injury. A decrease in the level of consciousness decline implies improved neurological function, whereas an increase denotes deterioration. A reduction in mortality (treatment outcome) reflects treatment success, while an elevation indicates failure. The remaining indicators are also important, but whether their increase or decrease is positive or negative depends on multifaceted factors, such as patient-specific conditions, disease stage, and admission timing. As shown in Table 1.

4 Numerical Experiments

4.1 Top-down descriptive norm optimization experiment

Although the v8/v9 guidelines reshaped the clinical usage propensities for key medical interventions such as hydroxychloroquine, azithromycin, and chloroquine, the Gaussian Mixture Models (GMMs) of these three medications exhibit identical mathematical structures and properties, differing only in numerical values. Here, we use the dynamic changes in azithromycin as a representative for experimental demonstration, while the others can be generalized using the same modeling framework.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_31

Table 1: PN = Prednisolone, MP = Methylprednisolone, ENX = Enoxaparin sodium. Significant causal effects from Double Machine Learning analysis with FDR correction (q-value < 0.01). Outcome: 1=death, 0=recovery. Consciousness decline: scale 0-2 with 2=severe decline (nearly always fatal).

Indicator	Control	Effect	Std Error	p-value	q-value	Lower CI	Upper CI	Signif
Outcome	MP	-0.0333	0.0079	2.31E-05	0.0001	-0.0488	-0.0179	***
Lymphocyte count	MP	1.1403	0.3210	0.0004	0.0012	0.5110	1.7695	**
Consciousness decline	MP	-0.0465	0.0151	0.0021	0.0077	-0.0761	-0.0168	**
Outcome	PN	-0.0540	0.0080	1.76E-11	1.94E-10	-0.0697	-0.0383	***
Lymphocyte count	PN	1.2646	0.3845	0.0010	0.0022	0.5110	2.0182	**
Consciousness decline	PN	-0.0892	0.0156	9.99E-09	1.10E-07	-0.1197	-0.0587	***
LDH level	ENX	54.1928	8.8018	7.41E-10	2.45E-08	36.9414	71.4443	***
Lymphocyte count	ENX	-0.6840	0.1938	0.0004	0.0012	-1.0639	-0.3042	**

Our opinion dynamics model operated on a bounded one-dimensional space $[-L, L]$ with $L = 2.0$, discretized into $N_x = 300$ points. The dynamics included diffusion ($d = 0.2$) and migration ($c = 1.0$) terms, initialized with uniform density $P_h = 1.0$. Three interpretable control parameters shaped the evolution: perception width (μ), perception scale (σ), and external potential strength ($k_{\text{potential}}$), each discretized into 15 values over clinically relevant ranges ($\mu, \sigma \in [0.05, 1.0]$; $k_{\text{potential}} \in [0.01, 3.0]$), yielding 3,375 possible actions.

The MDP state reduced the distribution to three sufficient statistics: mean (central tendency), standard deviation (spread), and Wasserstein-based similarity to target. We designed a composite reward function balancing multiple objectives with empirically determined weights: minimizing Wasserstein distance (-20.0), enhancing distribution similarity ($+10.0$), penalizing control effort ($-1.0 \times k_{\text{potential}}$), rewarding terminal convergence (up to $+50.0$), ensuring parameter smoothness (-0.002 per unit change), and preserving multimodality (up to $+25.0$ for correct peak count).

We implemented a dictionary-based Q-learning agent with adaptive state discretization. Training used $\alpha = 0.08$, $\gamma = 0.95$, and ϵ linearly decaying from 1.0 to 0.05 over $5,000$ episodes. We scheduled exploration-exploitation trade-off via linear decay to ensure thorough early exploration and focused later exploitation. A replay buffer (capacity $50,000$) trained on batches of 128 samples to break temporal correlations. We assessed policy performance through cumulative re-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_31

ward and Wasserstein distance, extracting optimal parameters after convergence for final distribution simulation.

Figure 2 presents a comprehensive comparison between reinforcement learning-optimized parameters and a random parameter baseline. The three panels evaluate performance across key dimensions: distribution alignment (Wasserstein distance), pattern fidelity (distribution similarity), and computational efficiency (convergence time). Each panel displays results from 100 randomly sampled parameter combinations drawn from clinically plausible ranges ($\mu, \sigma \in [0.05, 1.0]$; $k \in [0.01, 3.0]$), forming the boxplot distribution. The red dashed line and star marker indicate the performance of parameters discovered through our RL optimization framework. The green annotations quantify the percentage improvement achieved by the optimized parameters relative to the random parameter average. This visualization demonstrates that our approach consistently outperforms random parameter selection across all evaluation metrics, with particularly dramatic improvements in convergence speed and distribution alignment quality. The statistical significance of these improvements validates the robustness of our optimization methodology across the parameter space.

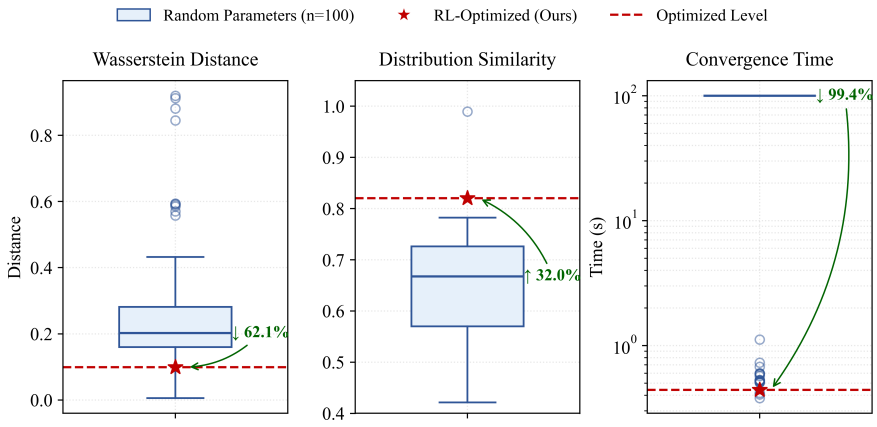

Fig. 2: Performance comparison between optimized parameters and 100 random parameter combinations across three evaluation metrics.

4.2 Bottom-up descriptive norm optimization

As previously mentioned in the dataset introduction, the bottom-up optimization of descriptive norms no longer requires computing the similarity between distributions in the opinion space and the target GMM. Instead, it transforms into a clinically-oriented mechanism.

Under the current MDP learning parameters, we select and evaluate the impact of a specific control column (e.g., prednisolone usage status control; pred-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_31

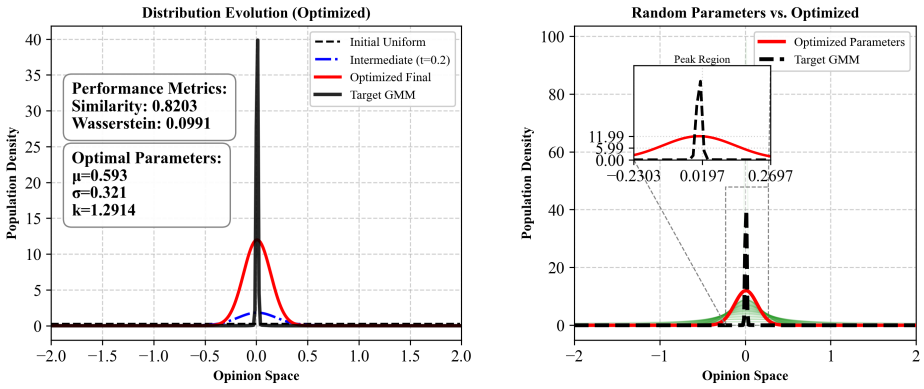

Fig. 3: Distribution evolution under optimized control parameters versus random parameter selections. Left panel shows progression from uniform initial state to optimized distribution matching target GMM (black). Right panel compares 100 random parameter trajectories (faint green) with optimized trajectory (red) against target distribution (black). Inset magnifies peak region (0.0197 ± 0.25).

nisolone, enoxaparin sodium, and methylprednisolone exhibit identical mathematical structures with only numerical differences, while other control columns can still be computed within this model) on core clinical indicators based on its causal effects. These indicators include mortality rate, absolute lymphocyte count throughout the disease course, and level of consciousness decline throughout the disease course. The reward function no longer pursues precise matching between the distribution and a fixed target, but instead evaluates whether the descriptive norm represented by the distribution generated under the current PDE parameters for the entire period will lead to positive medical outcomes.

The specific optimization goal is set as: not pursuing the maximization of indicators (to avoid clinically unrealistic extreme values), but ensuring that the predicted results are superior to the baseline levels from Period 4 dataset (mortality rate 0.0525, lymphocyte count 1.8394, consciousness decline 0.0944). Through this mechanism, we utilize MDP to train the kernel function parameters (μ and σ) and the external potential field strength k, thereby shaping a distribution evolution that guides practical tendencies toward clinical optimization.

The overall process is illustrated in Fig. 1: practice tendencies characterized by GMMs (historical data) $\rightarrow$ propensity mapper (physical meaning) $\rightarrow$ propensity mapping + Q-learning shaped PDE distribution (descriptive norm evolution) $\rightarrow$ practice indicator optimization.

The experimental procedure first extracts the overall usage propensity of prednisolone_stat_control from the GMM models of the five periods, as shown in Fig. 4. Specifically, for each period, we load the precomputed GMM result files, which contain the mixture weights π, mean matrix μ, and covariance matrix cov. Based on this, we construct a mapping function that encompasses all periods. Specifically, we calculate the minimum and maximum propensity values

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_31

(min_propensity and max_propensity, extended by 10 percent to accommodate variations), and design a serializable PropensityMapper class. This mapper transforms the center of gravity (weighted average position) of the PDE distribution $P(x)$ into a propensity value within the [0,1] range through linear normalization and S-shaped transformation (sigmoid function with $\sigma = 10$ to enhance sensitivity to extremes), ensuring the physical meaning of the mapping (left-skewed distributions correspond to conservative medication usage, right-skewed to aggressive medication usage, see Fig. 5).

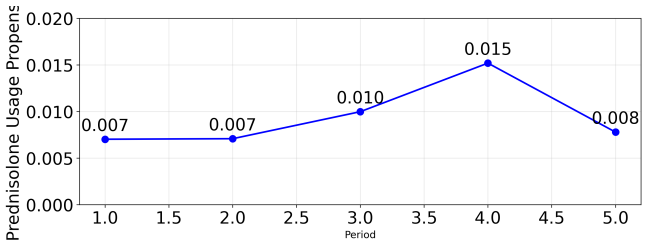

Fig. 4: Trend of prednisolone usage propensity across five clinical periods, peaking at period 4 with value 0.015.

Next, based on this mapping, we optimize randomly generated PDE distributions using Q-learning. The initial distribution is randomly generated: with 70% probability as a single Gaussian (center uniformly distributed in [-0.8, 0.8], width in [0.1, 1.0]), and with 30 % probability as multi-Gaussian distributions (2-5 peaks with random centers, widths, and weights). The spatial domain is set to $L = 1.0$ with $Nx = 200$ grid points; PDE parameters include diffusion coefficient $d = 0.05$, migration coefficient $c = 0.8$, and uniform density $P_h = 1.0$ to ensure numerical stability. The action space is discretized into 8×8×8=512 combinations ($\mu \in [0.1, 0.6]$, $\sigma \in [0.1, 0.6]$, $k \in [0.05, 1.2]$). The state space is simplified to 2 dimensions (distribution center of gravity and concentration, 1/standard deviation). During each step, solve_ivp (BDF method, atol=1e-4, rtol=1e-3, max_step=0.5) simulates PDE evolution up to $t = 50.0$.

The causal effects of prednisolone on core clinical indicators are shown in 1, where a one-unit increase in prednisolone usage propensity leads to a mortality rate reduction of -0.0540, an absolute lymphocyte count increase of +1.2646, and a consciousness decline reduction of -0.0892. Reward calculation is based on the mapping-derived propensity value, applying causal effects through the saturation effect ($1-e^{-5\cdot\text{propensity}}$), predicting clinical outcomes and comparing with Period 4 baselines: +10 reward per 0.01 unit improvement in mortality rate (capped at 5 times the baseline improvement), +8 reward per 0.5 unit improvement for lymphocyte count (capped at 3 times), +8 reward per 0.02 unit improvement for consciousness decline (capped at 3 times); if all indicators exceed baseline values by 2%, an additional +20 reward is granted; reality penalty is applied (e.g., -20 if mortality rate falls below 0.015); plus control cost penalty (action continuity

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_31

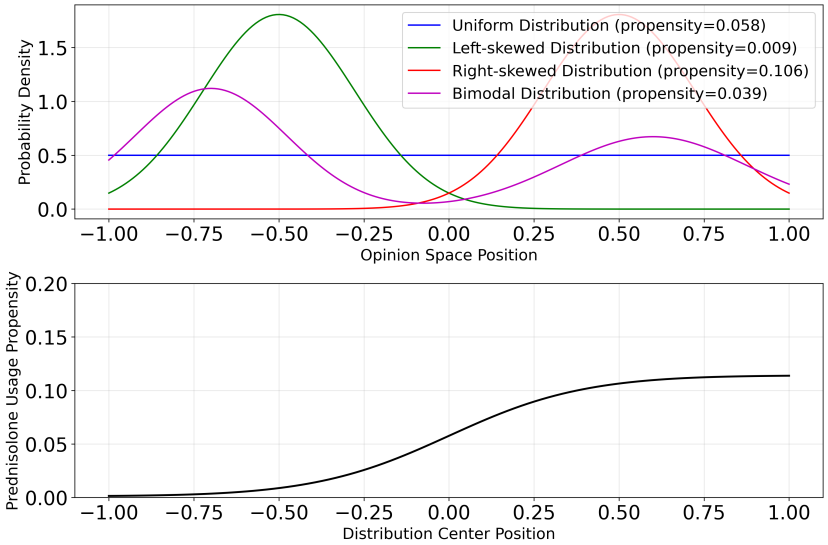

Fig. 5: The top subplot evaluates four representative opinion distributions, confirming that the mapped propensity accurately reflects macroscopic consensus. The bottom subplot visualizes the continuous Sigmoid transfer curve, ensuring a smooth, bounded mapping from opinion centers to clinical decisions.

penalty of 0.1 times the squared difference between consecutive actions). The Q-learning agent employs a learning rate $\alpha = 0.08$, discount factor $\gamma = 0.95$, initial exploration rate $\epsilon = 1.0$ (exponentially decayed to 0.05), state discretization (20 bins for center of gravity, 20 bins for concentration), and incorporates experience replay (buffer size 10,000, batch size 64) to enhance training stability. Training runs for 2,000 episodes, with clinical history and parameters recorded every 50 episodes, reporting average reward and estimated remaining time. The experimental results are presented in Fig. 6, Fig. 7, and Fig. 8.

5 Discussion of experimental results

As shown in the top panel of Fig. 5, the right-skewed distribution representing aggressive treatment in the propensity mapper has a propensity value of 0.1062. As demonstrated in Fig. 6 and Fig. 7, our optimized emergent collective strategy exceeds the baseline of Period 4 in the dataset and produces a right-skewed distribution consistent with the mapper's prediction. The interpretability of our results aligns with the PDE mathematical model inferences and also matches the structural properties of the dataset.

From Fig. 8, it can be seen that prednisolone's improvement on the three clinical indicators based on causal effects significantly exceeds the empirical practice

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_31

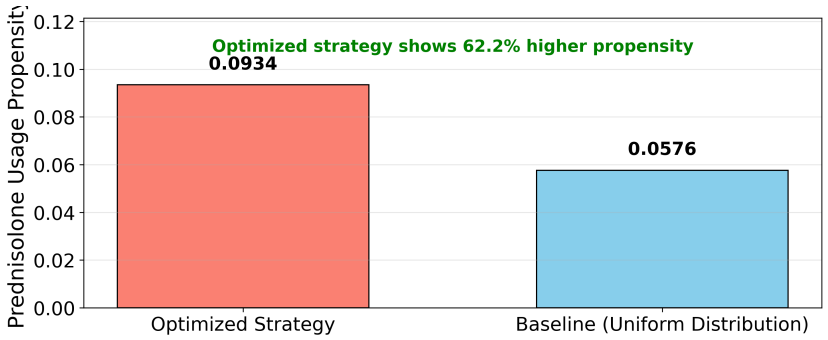

Fig. 6: Optimized strategy achieves 0.0934 prednisolone usage propensity, 62.2% higher than the baseline uniform distribution (0.0576).

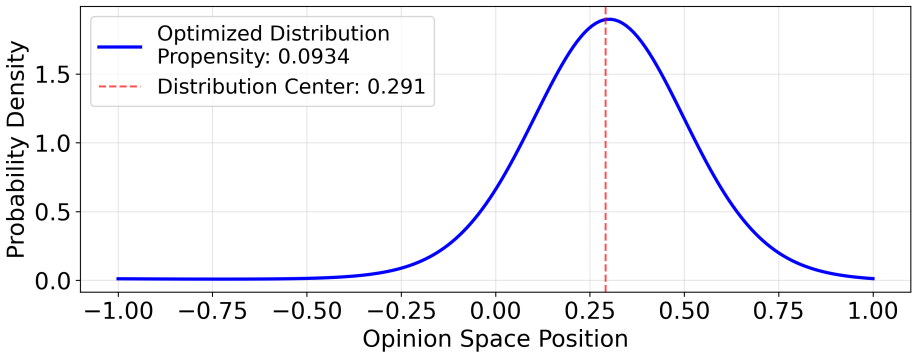

Fig. 7: Optimized prednisolone usage opinion distribution with peak at position 0.291, corresponding to 0.0934 usage propensity.

patterns of healthcare professionals in Period 4 of the dataset, demonstrating greater effectiveness when facing emergent new variants.

Looking back at our top-down experiment, since continuous-action reinforcement learning frameworks such as SAC (Soft Actor-Critic) and PPO (Proximal Policy Optimization) were not employed in this work, the left panel of Figure 3 demonstrates that while the optimized distribution achieves a similarity score of 0.8203 with a Wasserstein distance of 0.0991 relative to the target GMM, it fails to capture the prominent peak characteristic of azithromycin in the original dataset. The right panel of Figure 3 provides a zoomed-in view comparing the optimized distribution against 100 random parameter combinations, showing that the optimized solution maintains a closer shape alignment with the target GMM than random parameterizations across the clinically relevant opinion range.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_31

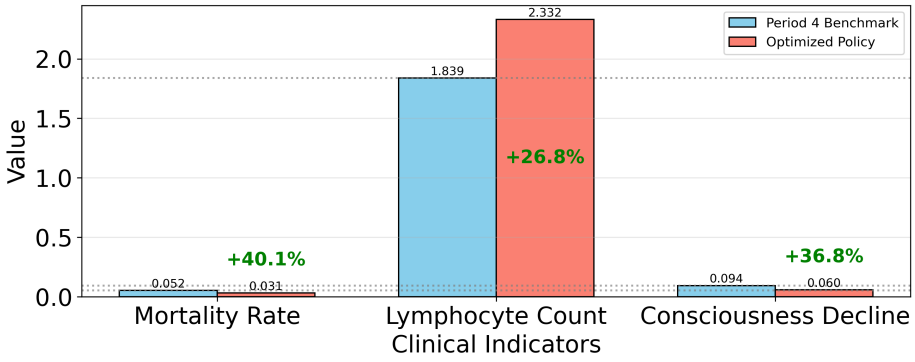

Fig. 8: Optimized policy demonstrates improved clinical outcomes with 40.1% lower mortality rate, 26.8% higher lymphocyte count, and 36.8% reduced consciousness decline compared to Period 4 benchmark.

6 Conclusion and Future Work

In the top-down experiment, we did not implement PPO or other continuous-action reinforcement learning frameworks. While these approaches could be valuable for future work requiring more precise fitting of target norms—particularly under stringent macro-level normative constraints—their limited interpretability remains a concern. The focus of this paper, however, centers on demonstrating the feasibility of severely optimized norm models rather than pursuing maximal representational accuracy.

Both experiments presented here employ a parameter transfer methodology: first optimizing parameters through the MDP-PDE framework, then applying these optimized parameters to guide norm evolution. Intriguingly, a promising future direction involves enabling agents to autonomously explore the normative landscape—first learning existing norms, then deliberately violating them to facilitate the emergence of new collective norms—a paradigm that could significantly advance our understanding of bottom-up normative evolution in complex adaptive systems.

Acknowledgments. The research was supported by the Russian Science Foundation, agreement No. 24-11-00272, https://rscf.ru/project/24-11-00272/.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Austin, P.C.: Assessing covariate balance when using the generalized propensity score with quantitative or continuous exposures. Statistical methods in medical research **28**(5), 1365–1377 (2019)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_31

2. Bakshy, E., Messing, S., Adamic, L.A.: Exposure to ideologically diverse news and opinion on facebook. Science **348**(6239), 1130–1132 (2015)
3. Boi, S., Capasso, V., Morale, D.: Modeling the aggregative behavior of ants of the species polyergus rufescens. Nonlinear Analysis: Real World Applications **1**(1), 163–176 (2000)
4. Brown, D.W., Greene, T.J., Swartz, M.D., Wilkinson, A.V., DeSantis, S.M.: Propensity score stratification methods for continuous treatments. Statistics in medicine **40**(5), 1189–1203 (2021)
5. Chen, Y., Bosse, T., Woensdregt, M.: Social norms as an interactive process: An agent-based cognitive modelling study. In: Proceedings of the Annual Meeting of the Cognitive Science Society. vol. 46 (2024)
6. Di Francesco, M., Fagioli, S.: Measure solutions for non-local interaction pdes with two species. Nonlinearity **26**(10), 2777 (2013)
7. Dickinson, R., Makowski, D., van Marwijk, H., Ford, E.: Exploring the role of news outlets in the rise of a conspiracy theory: Hydroxychloroquine in the early days of covid-19. COVID **4**(12), 1873–1896 (2024)
8. Hillen, T., Painter, K.J.: A user's guide to pde models for chemotaxis. Journal of mathematical biology **58**(1), 183–217 (2009)
9. Horstmann, D.: From 1970 until present: the keller-segel model in chemotaxis and its consequences (2003)
10. Huber, M., Hsu, Y.C., Lee, Y.Y., Lettry, L.: Direct and indirect effects of continuous treatments based on generalized propensity score weighting. Journal of Applied Econometrics **35**(7), 814–840 (2020)
11. Keller, E.F., Segel, L.A.: Initiation of slime mold aggregation viewed as an instability. Journal of theoretical biology **26**(3), 399–415 (1970)
12. Kwon, J., Zhi-Xuan, T., Tenenbaum, J., Levine, S.: When it is not out of line to get out of line: The role of universalization and outcome-based reasoning in rule-breaking judgments (2023)
13. Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J., Cushman, F.: The logic of universalization guides moral judgment. Proceedings of the National Academy of Sciences **117**(42), 26158–26169 (2020)
14. Li, C., Derevitskii, I., Kovalchuk, S.: Modeling descriptive norms in multi-agent systems: An auto-aggregation pde framework with adaptive perception kernels (2026), https://arxiv.org/abs/2601.06557
15. Li, C., Petruchik, O., Grishanina, E., Kovalchuk, S.: Multi-agent norm perception and induction in distributed healthcare. Journal of Biomedical Informatics p. 104835 (2025)
16. Littman, M.L.: Markov games as a framework for multi-agent reinforcement learning. In: Machine learning proceedings 1994, pp. 157–163. Elsevier (1994)
17. Mogilner, A., Edelstein-Keshet, L.: A non-local model for a swarm. Journal of mathematical biology **38**(6), 534–570 (1999)
18. Saag, M.S.: Misguided use of hydroxychloroquine for covid-19: the infusion of politics into science. Jama **324**(21), 2161–2162 (2020)
19. Sayama, H.: Enhanced ability of information gathering may intensify disagreement among groups. Physical Review E **102**(1), 012303 (2020)
20. Yazdany, J., Kim, A.H.: Use of hydroxychloroquine and chloroquine during the covid-19 pandemic: what every clinician should know (2020)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_31