Research Highlight: Ambiguous Dynamic Treatment Regimes: A Reinforcement Learning Approach

by SOROUSH SAGHAFIAN
HARVARD UNIVERSITY, CAMBRIDGE MA 02138
SOROUSH_SAGHAFIAN@HKS.HARVARD.EDU

1 Introduction

A central objective in precision medicine, public policy, and economics is to discover *when* and *how* to dynamically make decisions affecting individuals—personalized to their evolving characteristics—so as to maximize outcomes over time. Dynamic Treatment Regimes (DTRs) formalize this objective: a DTR is a sequence of decision rules that maps a subject's observable history to a recommended action at each decision epoch [12, 22]. When learning and optimizing them using observational data, however, DTRs require *sequential ignorability*—the assumption that, conditional on measured covariates, treatment assignment is independent of all future potential outcomes [13, 14]. In various real-world applications, this assumption is almost never verifiable and is routinely violated. In healthcare, for example, unmeasured disease severity, patient adherence, or physician intuition can influence both the treatment selected and the subsequent outcome.

The challenge is compounded when unmeasured confounders are *time-varying*, meaning that they could be affected by previous decisions. In such settings, naïve adjustment methods that do not account for this feedback create additional bias [15]. Even if one postulates a specific causal model for the dynamics of unobserved confounders and their impact on observed data— the data generating model—such a model is itself subject to substantial ambiguity, and hence, any useful data-driven learning algorithm should be able to handle *model ambiguity* [see, also [19, 21]].

Saghafian [17] addresses both challenges simultaneously. The paper makes four inter-related contributions that together form a coherent and practically important framework for learning suitable dynamic decisions from data:

1. It extends DTRs to *Ambiguous* DTRs (ADTRs) in which the causal impact of any treatment policy is evaluated over a *cloud* of plausible data-generating models rather than a single assumed model.
2. It connects ADTRs to Ambiguous Partially Observable Markov Decision Processes (APOMDPs) [16], reinterpreting time-varying unobserved confounders as latent states with ambiguous transition dynamics.
3. It develops two efficient off-policy Reinforcement Learning (RL) algorithms—*Direct Augmented V-Learning* (DAV-Learning) and *Safe Augmented V-Learning* (SAV-Learning)—that learn optimal treatment regimes from observational data under model ambiguity.
4. It introduces the concept of *two-way personalization*: the resulting treatment policies are tailored simultaneously to the individuals' (e.g., patients) varying context variables and to the decision-maker's (e.g., a physician's) behavioral attitude toward ambiguity, parameterized by a pessimism level $\alpha \in [0, 1]$.

The advantages of the proposed framework in Saghafian [17] is investigated via a case study of New Onset Diabetes After Transplantation (NODAT), a condition that arises from the diabetogenic side-effect of immunosuppressive drugs (e.g., tacrolimus) given to organ transplant recipients [1, 2, 8, 9, 10]. Physicians must jointly manage the risk of organ rejection (requiring high tacrolimus) and the risk of diabetes onset (requiring low tacrolimus and possibly insulin). Clinical data on 407 kidney transplant patients (over 63,000 observations) collected at the Mayo Clinic provide the empirical testbed.

2 The ADTR Framework

Let $(O_t, A_t)_{t\in\mathcal{T}}$ denote the observed covariates and actions at each decision epoch. Let S_t denote the unobserved (latent) health state at time t. A treatment regime $\lambda = (\lambda_t)_{t\in\mathcal{T}}$ maps the observable history $H_t^o = (O_1, A_1, \ldots, O_t)$ to a probability distribution over actions. The overall gain under λ is the discounted

sum of immediate gains $G_t = g(S_t, A_t)$:

$$\Gamma_T(\lambda) = \sum_{t \in \mathcal{T}} \beta^{t-1} G_t^{\lambda}, \quad \beta \in (0,1).$$

In a standard DTR, one estimates the optimal λ^* by maximizing $\mathbb{E}[\Gamma_T(\lambda)]$ over the observed data, implicitly committing to a single causal model, and hence, a unique imposed distribution for $\Gamma_T(\lambda)$. In ADTRs [17], the causal model is not uniquely identified from the observed data. The set $\mathcal{M}$ of all models consistent with the observations constitutes the *ambiguity set*. Under each model $m \in \mathcal{M}$, the gain $\Gamma_T(\lambda)$ considered as a potential outcome variable of interest $Y(\lambda) = \Gamma_T(\lambda)$ follows a different distribution f^m. To compare policies, thus, Saghafian [17] adopts the α-Maximin Expected Utility (α-MEU) criterion [5, 6, 16]:

$$\mathrm{MEU}_\alpha[Y(\lambda)] = \alpha \inf_{f^m \in \hat{\mathcal{F}}} \mathbb{E}_{f^m}[Y(\lambda)] + (1-\alpha) \sup_{f^m \in \hat{\mathcal{F}}} \mathbb{E}_{f^m}[Y(\lambda)], \quad \alpha \in [0,1].$$

Here $\alpha = 1$ recovers the classical worst-case (maximin) criterion, $\alpha = 0$ is the optimistic (maximax) criterion, and intermediate values interpolate between the two. When $|\mathcal{M}| = 1$, MEU_α reduces to the standard expectation, so the traditional DTR optimality criterion is a special case.

This formulation is important for two reasons. First, it avoids the well-documented over-conservatism of the pure maximin view, which Savage (1951) described as "ultrapessimistic" [17]. Second, and crucially for real-world applications, it allows the decision-maker's own attitude towards ambiguity to be encoded in α, yielding one the paper's central conceptual contribution: *two-way personalization*. A decision-maker (e.g., a physician) who is particularly averse to worst-case outcomes (high α) will receive different recommendations than one who is more optimistic (low α). When needed, however, α can be viewed as a tuning parameter and optimized to gain the best overall recommendation.

Connection to APOMDPs. When the dynamics of the state variables satisfy the Markov property, the ADTR problem maps cleanly onto an Ambiguous Partially Observable Markov Decision Process (APOMDP) [16]. In this mapping, the unobserved confounders S_t become the latent state of the POMDP, while the observable history H_t^o is used to form a Bayesian belief distribution π_t over S_t via a standard belief-updating operator [16, 17]. The ambiguity set $\mathcal{M}$ translates into multiple candidate transition and emission probability matrices (P_m^a, Q_m^a). Crucially, Saghafian [16] has shown that the APOMDP value function $V_t(\pi)$ under some conditions is *piecewise linear and continuous* in the belief π—a structural property that the paper exploits to design computationally tractable data-driven learning algorithms.

Generalized Sequential Importance Sampling. Before turning to the Markovian case, Saghafian [17] establishes a non-Markovian baseline approach termed *Generalized Sequential Importance Sampling* (GSIS). GSIS reweights observed trajectories by importance sampling weights $w_t(\lambda^e) = \frac{\lambda_t^e(A_t|H_t^o)}{\lambda_t^b(A_t|H_t^o)}$, where λ^b is the behavior (data-generating) policy and λ^e is any policy to be evaluated. Under sequential ignorability and almost-sure overlap between evaluation and behavior policies, GSIS yields an MEU_α-unbiased estimator of $\Gamma_T(\lambda^e)$. The paper also extends GSIS to the practically important case of *Bounded Unobservable Confounding* (BUC), where the sequential ignorability no longer holds, but unobserved confounders affect treatment propensities only within known multiplicative bounds $\eta_t^m \in [1, \infty)$, providing approximate unbiased estimation even when sequential ignorability fails.

Finally, under Markovian behavior, [17] demonstrates that causal relations in observed ADTR data can be represented via a directed acyclic graph (DAG) depicted in Fig. 1, which is a DAG representation of APOMDPs. This enables developing causal RL algorithms as discussed in the next section.

3 Theoretical Results

Weight-Adjusted Bellman Equations. A central technical result in Saghafian 17 is based on a *weight-adjusted Bellman equation* for the value function under an evaluation policy μ^e in any single POMDP model m:

$$V_{T-t+1}^{m,\mu^e}(\pi_t) = \mathbb{E}_m \left[\frac{\mu_t^e(A_t|\Pi_t^m)}{\mu_t^b(A_t|\Pi_t^m)} \Big(G_t + \beta V_{T-t}^{m,\mu^e}(T(\Pi_t^m, A_t, O_t, m)) \Big) \,\Big|\, \Pi_t^m = \pi_t \right].$$

This equation—which holds whenever the behavior policy satisfies positivity and sequential ignorability after conditioning on belief states—is fundamental: it expresses the value of any evaluation policy purely in terms of observable quantities, enabling estimation from data.

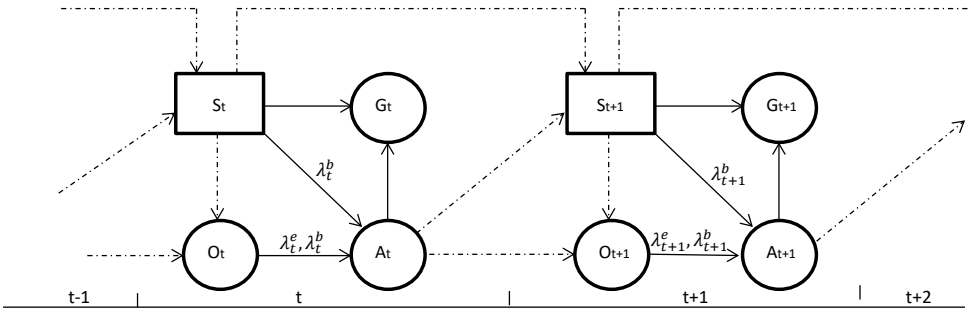

Figure 1: DAG representation of APOMDPs (Fig. 1 in [17]). *Circles:* observable variables; *Rectangles:* unobservable variables; *Solid arrows:* unambiguous causal mechanisms; *Dashed arrows:* ambiguous causal mechanisms.

Under BUC, the equation generalizes to sandwich bounds on the value function using weight modifiers described in [17] that incorporate the confounding bounds η_t^m. This BUC extension underpins the DAV-Learning-BUC and SAV-Learning-BUC variants of the main algorithms.

Reinforcement Learning Algorithms: DAV-Learning and SAV-Learning. Both algorithms learn the value function in the class $\mathcal{V}$ of piecewise linear and continuous functions on the belief simplex Δ_S, using a parametric form $V_\infty^{m,\mu^e}(\pi; c) = (b(\pi))'c$ where $b(\pi)$ is a predefined basis (c is used here in place of ψ in [17] for the ease of notation). The algorithms differ in *when* they incorporate model ambiguity.

DAV-Learning (Algorithm 1 in [17]) first estimates the value function separately for each model $m \in \mathcal{M}$ via regularized minimization of the empirical Bellman residual:

$$\hat{c}_n^{m,\mu^e} = \arg\min_{c \in \mathcal{C}} \big\{ (\varphi_n^{m,\mu^e}(c))' \Omega \, \varphi_n^{m,\mu^e}(c) + \theta_n P(c) \big\},$$

and then combines model-specific gains via the α-MEU operator at the end of the horizon:

$$\hat{\Gamma}_\infty(\mu^e) = \alpha \inf_{m \in \mathcal{M}} \hat{\Gamma}_\infty^m(\mu^e) + (1-\alpha) \sup_{m \in \mathcal{M}} \hat{\Gamma}_\infty^m(\mu^e).$$

SAV-Learning (Algorithm 2 in [17]) instead incorporates ambiguity *up front* by estimating the value function parameter as the α-MEU of the individual model estimates:

$$\hat{c}_n^{\mu^e} = \alpha \, \hat{c}_n^{\underline{m},\mu^e} + (1-\alpha) \, \hat{c}_n^{\overline{m},\mu^e},$$

where $\underline{m}$ and $\overline{m}$ minimize and maximize the parameter norm over $\mathcal{M}$. This "safe" integration of ambiguity at the estimation stage makes the resulting policy more robust to the choice of α, at the cost of some mean performance relative to DAV-Learning.

Asymptotic Properties of the Algorithms. The paper establishes the asymptotic behavior of both algorithms under five regularity conditions covering the parameter space, the policy space, the trajectory process, and the model space. The key results (Theorems 1 and 2 in [17]) are:

1. **Weak consistency**: For any fixed evaluation policy μ^e and model m, $\hat{c}_n^{m,\mu^e} \xrightarrow{p} c_*^{m,\mu^e}$ and $\hat{\Gamma}_\infty(\mu^e) \xrightarrow{p} \Gamma_\infty(\mu^e)$.
2. **Asymptotic normality**: $\sqrt{n}(\hat{c}_n^{m,\mu^e} - c_*^{m,\mu^e})$ converges in distribution to a zero-mean Gaussian process in $\ell^\infty(\mathcal{M})$.
3. **Consistency of the optimal policy**: $d_{\mathcal{M}}(\hat{\mu}_n^{e*}, \mu^{e*}) \xrightarrow{p} 0$ and $\hat{\Gamma}_\infty(\hat{\mu}_n^{e*}) \xrightarrow{p} \Gamma_\infty(\mu^{e*})$.

A notable technical feature is that the proofs must handle both non-i.i.d. trajectory data (which are absolutely regular stationary processes with β-mixing coefficients satisfying a summability condition) and the presence of multiple ambiguous models. The asymptotic results in [17] leverage findings from *empirical process theory* for dependent data [3, 7] to establish *Donsker properties* in $\ell^\infty(\mathcal{M})$.

The tuning parameter θ_n must satisfy $\theta_n = o_p(n^{-1/2})$, which is the standard rate for penalized estimators that achieve consistency and asymptotic normality simultaneously. These results guarantee that with enough data, ADTRs can be reliably estimated and that uncertainty in the estimated optimal policy diminishes at the usual $\sqrt{n}$ rate.

4 Empirical Results

Case Study: NODAT at Mayo Clinic. The clinical data set comprises 407 kidney transplant recipients observed monthly for 12 months post-transplant (13 covariates, 9 latent health states, 4 possible actions combining tacrolimus dosing and insulin use). After cubic spline interpolation to fill data gaps, the full data set contains 63,492 observations. The ambiguity set $\mathcal{M}$ contains 4 models constructed via an entropy ball around Baum–Welch estimates of the POMDP emission and transition matrices.

Table 5 of Saghafian [17] reports total discounted gains (discount factor $\beta = 0.95$) under the observed clinical regime and the two proposed algorithms across five values of $\alpha \in \{0.00, 0.25, 0.50, 0.75, 1.00\}$. The main findings are:

- **Both algorithms substantially outperform observed clinical practice.** DAV-Learning improves over the observed regime by 10%–42% and SAV-Learning by 10%–32%, depending on α.
- **DAV-Learning achieves higher mean gains**, particularly for low α (optimistic) settings. At $\alpha = 0$, DAV-Learning yields a mean total gain of 2.085 versus 1.472 under the observed regime.
- **SAV-Learning is far more robust to the choice of α.** Whereas DAV-Learning performance degrades markedly as α increases from 0 to 1 (from 2.085 to 1.609), SAV-Learning spans a narrower range (1.949 to 1.606). An optimizer using SAV-Learning need not carefully tune α.

These findings reflect the structural difference between the algorithms: SAV-Learning's "safe" up-front aggregation of model ambiguity effectively buffers against the choice of pessimism level, while DAV-Learning's end-of-horizon aggregation preserves sensitivity to α but also unlocks higher upside for optimistic decision-makers.

Synthetic Data Experiments. Synthetic experiments simulate 100 patients over 10 periods under $|\mathcal{M}| = 10$ misspecified models generated from Dirichlet distributions with random parameters, with a known true model for benchmarking. Key results (Table 6 in [17]) mirror the clinical findings: both DAV-Learning and SAV-Learning improve over the observed regime, and SAV-Learning's performance is more stable across α values.

Robustness to Model Ambiguity. The most striking empirical result concerns robustness. The paper benchmarks all algorithms against an imaginary oracle who knows both the true data-generating model and the optimal policy under that model. The *gain loss* (regret relative to the oracle) exhibits a U-shaped curve in α: both extreme pessimism ($\alpha = 1$) and extreme optimism ($\alpha = 0$) are suboptimal choices, while intermediate values (around $\alpha \approx 0.25$) minimize regret. More importantly, the maximum gain loss across all values of α for all four algorithms remains below 0.6%. This is a remarkable finding: a decision-maker operating under complete uncertainty about the data-generating process can, by using DAV-Learning or SAV-Learning, achieve performance essentially indistinguishable from an oracle who knows the truth.

This result in [17] provides concrete empirical support for the paper's philosophical position: a "cloud of models" approach to causal inference—taking a middle ground between fully model-based and fully model-free reasoning—can simultaneously deliver personalization, interpretability, and robustness.

5 Conclusions

The core insight in Saghafian [17]—that model ambiguity is not merely a nuisance to be minimized but an inherent feature of observational data that should be *built into* the learning framework—makes a significant and timely contribution to the intersection of causal inference, reinforcement learning, and sequential decision-making. It can reorients how researchers might think about off-policy evaluation under unobserved confounding. The paper's "cloud of models" methodology explicitly positions ADTRs between fully structural causal models (which commit to a single data-generating process) and purely nonparametric, model-free methods (which make no structural assumptions). This middle ground echoes calls by Manski [11] and others for causal inference methods that remain valid across the set of all feasible models.

References

[1] Boloori, A., Saghafian, S., Chakkera, H. A., and Cook, C. B. (2015). Characterization of remitting and relapsing hyperglycemia in post-renal-transplant recipients. *PLoS One,* 10(11), e0142363.

[2] Boloori, A., Saghafian, S., Chakkera, H. A., and Cook, C. B. (2020). Data-driven management of post-transplant medications: An ambiguous partially observable Markov decision process approach. *Manufacturing and Service Operations Management*, 22(5), 1066-1087.

[3] Dedecker, J., and Louhichi, S. (2002). Maximal inequalities and empirical central limit theorems. In *Empirical Process Techniques for Dependent Data*, pp. 137–159. Birkhäuser, Boston.

[4] Frank, R. G., and Zeckhauser, R. J. (2007). Custom-made vs. ready-to-wear treatments: Behavioral propensities in physicians' choices. *Journal of Health Economics*, 26(6):1101–1127.

[5] Ghiradato, P., Maccheroni, F., and Marinacci, M. (2004). Differentiating ambiguity and ambiguity attitude. *Journal of Economic Theory*, 118:133–173.

[6] Hurwicz, L. (1951). Optimality criteria for decision making under ignorance. Cowles Commission Discussion Paper: Statistics No. 370.

[7] Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York.

[8] Munshi, V. N., Saghafian, S., Cook, C. B., Steidley, D. E., Hardaway, B., and Chakkera, H. A. (2020). Incidence, risk factors, and trends for postheart transplantation diabetes mellitus. *The American Journal of Cardiology*, 125(3), 436-440.

[9] Munshi, V. N., Saghafian, S., Cook, C. B., Werner, K. T., and Chakkera, H. A. (2020). Comparison of post-transplantation diabetes mellitus incidence and risk factors between kidney and liver transplantation patients. *PloS one*, 15(1), e0226873.

[10] Munshi, V. N., Saghafian, S., Cook, C. B., Aradhyula, S. V., and Chakkera, H. A. (2021). Use of imputation and decision modeling to improve diagnosis and management of patients at risk for new-onset diabetes after transplantation. *Annals of Transplantation*, 26, e928624-1.

[11] Manski, C. F. (2021). Econometrics for decision making: Building foundations sketched by Haavelmo and Wald. *Econometrica*, 89(6):2827–2853.

[12] Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B*, 65(2):331–355.

[13] Murphy, S. A., van der Laan, M. J., Robins, J. M., and CPPRG (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423.

[14] Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period. *Mathematical Modelling*, 7(9–12):1393–1512.

[15] Robins, J., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

[16] Saghafian, S. (2018). Ambiguous partially observable Markov decision processes: Structural results and applications. *Journal of Economic Theory*, 178:1–35.

[17] Saghafian, S. (2024). Ambiguous dynamic treatment regimes: A reinforcement learning approach. *Management Science*, 70(9):5667–5690.

[18] Saghafian, S., and Murphy, S. A. (2021). Innovative healthcare delivery: The scientific and regulatory challenges in designing mHealth interventions. *NAM Perspectives*, National Academy of Medicine, Washington, DC.

[19] Saghafian, S., and Tomlin, B. (2016). The newsvendor under demand ambiguity: Combining data with moment and tail information. *Operations Research*, 64(1), 167-185.

[20] Saghafian, S., Tomlin, B., and Biller, S. (2022). The Internet of Things and information fusion: Who talks to who? *Manufacturing & Service Operations Management*, 24(1):333–351.

[21] Saghafian, S. (2025). *Insight-driven Problem Solving: Analytics Science to Improve the World*. Cambridge University Press.

[22] Tsiatis, A. A., Davidian, M., Holloway, S. T., Laber, E. B., and Kosorok, M. R. (2019). *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. Chapman and Hall/CRC, Boca Raton, FL.