

IA Pública Segura: Causalidad, Robustez y Decisión

Safe Public AI: Causality, Robustness, and Decision

Ricardo Alonzo Fernández Salguero ¹

¹ Universidad Politécnica de Cataluña.

ricardo.alonzo.fernandez@upc.edu (Autor de correspondencia)

Resumen. Este trabajo desarrolla una formalización matemática integrada para sistemas de inteligencia artificial orientados a la optimización de políticas públicas bajo restricciones institucionales, presupuestarias, causales y distributivas. El enfoque propuesto articula procesos de decisión de Markov restringidos, medidas ocupacionales, dualidad, condiciones KKT, inferencia causal secuencial, evaluación fuera de política, robustez ante incertidumbre y criterios de seguridad respecto a políticas base. La contribución central consiste en modelar a la IA pública no como un predictor aislado, sino como un operador compuesto que transforma datos observacionales, parámetros normativos y restricciones auditables en una política factible y matemáticamente justificable. A lo largo del texto se demuestran resultados de existencia, realizabilidad de medidas ocupacionales, dualidad fuerte, mejora de política, diferencias de desempeño, sensibilidad ante errores de especificación, contracción robusta bajo rectangularidad y estabilidad local del selector multiobjetivo. Asimismo, se formalizan restricciones de equidad, implementabilidad y estabilidad institucional, mostrando que la optimización pública exige distinguir cuidadosamente entre predicción, causalidad, factibilidad normativa y legitimidad política. El resultado es una arquitectura analítica auditable y extensible para el diseño de sistemas de apoyo a la decisión pública que combinan eficiencia algorítmica con seguridad, robustez y responsabilidad institucional.

Palabras clave: IA; políticas públicas; MDP; medidas ocupacionales; dualidad; causalidad; robustez; seguridad; equidad; evaluación contrafactual

Abstract. This paper develops an integrated mathematical formalization for artificial intelligence systems oriented toward optimizing public policies under institutional, budgetary, causal, and distributive constraints. The proposed approach articulates constrained Markov decision processes, occupancy measures, duality, KKT conditions, sequential causal inference, off-policy evaluation, robustness under uncertainty, and safety criteria with respect to baseline policies. The central contribution consists in modeling public AI not as an isolated predictor, but as a composite operator that transforms observational data, normative parameters, and auditable constraints into a feasible and mathematically justifiable policy. Throughout the text, results are proved on existence, realizability of occupancy measures, strong duality, policy improvement, performance differences, sensitivity to specification errors, robust contraction under rectangularity, and local stability of the multiobjective selector. Equity, implementability, and institutional stability constraints are also formalized, showing that public optimization requires carefully distinguishing among prediction, causality, normative feasibility, and political legitimacy. The result is an auditable and extensible analytical architecture for designing public decision-support systems that combine algorithmic efficiency with safety, robustness, and institutional responsibility.

Keywords: AI; public policy; MDP; occupancy measures; duality; causality; robustness; safety; equity; counterfactual evaluation

Historial: Recibido: 28 de marzo de 2026 | Aceptado: 10 de mayo de 2026

Ética/Financiamiento: El autor declara no tener conflictos de interés. La investigación fue autofinanciada y no recibió financiamiento externo.

Este documento está bajo una licencia
CC BY-NC-ND 4.0.

Cómo citar (APA): Fernández Salguero, R. A. (2026). IA Pública Segura: Causalidad, Robustez y Decisión. *Conocimiento i Política*, 7(1), 93-161. DOI: 10.64480/cip.2026.n7.03.

1. Introducción

La incorporación de inteligencia artificial en la formulación, evaluación y ejecución de políticas públicas constituye una de las transformaciones más significativas de la administración contemporánea. Sin embargo, el uso de algoritmos en este ámbito plantea una exigencia mucho más severa que en contextos comerciales o puramente predictivos. En el sector público no basta con maximizar precisión empírica o rendimiento promedio: toda recomendación algorítmica debe ser compatible con restricciones presupuestarias, estándares de equidad, límites institucionales, criterios de seguridad, trazabilidad causal y exigencias de legitimidad democrática. Dicho de otro modo, una política sugerida por una IA puede ser estadísticamente eficiente y, al mismo tiempo, normativamente inaceptable, causalmente infundada o institucionalmente inviable.

Esta tensión obliga a abandonar la idea de la IA como un simple instrumento de predicción y a reconceptualizarla como un sistema formal de decisión bajo restricciones. En ese marco, los procesos de decisión de Markov restringidos ofrecen una base matemática especialmente adecuada, porque permiten representar dinámicas intertemporales, recompensas sociales, costos, restricciones y trayectorias de decisión dentro de una misma estructura analítica. No obstante, la optimización dinámica por sí sola resulta insuficiente. Si los parámetros del modelo han sido inferidos de datos observacionales sin una capa causal rigurosa, el algoritmo puede optimizar asociaciones espurias. Si el sistema no incorpora robustez frente a incertidumbre, puede colapsar ante shocks o cambios estructurales. Si no se modelan equidad, capacidad administrativa y estabilidad institucional, la solución óptima en el papel puede fracasar en la implementación real.

El propósito de este trabajo es desarrollar una formalización matemática integrada de una IA para políticas públicas que incorpore, de manera explícita y articulada, cinco dimensiones fundamentales: optimización dinámica restringida, identificación causal, evaluación fuera de política, robustez ante incertidumbre y seguridad normativa respecto a una política base. Para ello, el documento construye una arquitectura en capas. La primera capa, causal-estadística, transforma datos en estimaciones de transiciones, recompensas y restricciones bajo supuestos identificables. La segunda capa, de optimización, resuelve un problema dinámico restringido mediante medidas ocupacionales, dualidad y condiciones de optimalidad. La tercera capa, normativa-institucional,

introduce pesos sociales, umbrales distributivos, restricciones de implementabilidad y reglas de despliegue seguro. Esta separación no es meramente expositiva, sino analíticamente decisiva, porque permite distinguir con precisión entre error de modelo, error de optimización y error de transferencia al mundo real.

La relevancia de este enfoque radica en que la política pública es, por naturaleza, un problema secuencial, incierto y normativamente cargado. Las decisiones actuales alteran estados futuros, distribuciones de recursos, incentivos de los agentes y condiciones de factibilidad institucional. Por ello, el diseño de una IA pública exige una teoría que conecte dinámica intertemporal, inferencia causal y restricciones sociales dentro de un mismo marco matemático. Este trabajo responde a esa necesidad mostrando que es posible construir una base formal en la que la eficiencia algorítmica no sea confundida con legitimidad normativa, ni la correlación empírica con causalidad, ni la optimalidad en un modelo estimado con seguridad en el despliegue real.

La tesis general que guía el texto es que una IA pública adecuada debe concebirse como un operador auditable que transforma información empírica y criterios normativos explícitos en políticas factibles, robustas y justificables. Desde esta perspectiva, el objetivo no es automatizar ciegamente la decisión estatal, sino formalizar bajo qué condiciones una recomendación algorítmica puede ser considerada racional, segura y compatible con los fines del gobierno democrático. Así, el documento no sólo desarrolla herramientas matemáticas, sino que también delimita sus supuestos, su alcance y sus límites estructurales, con el fin de contribuir a una teoría rigurosa del uso responsable de la inteligencia artificial en la acción pública.

2. Revisión de la literatura

La integración de la Inteligencia Artificial y el Aprendizaje por Refuerzo, o “Reinforcement Learning” (RL), en el diseño y la ejecución de políticas públicas representa una de las fronteras más prometedoras y matemáticamente desafiantes de la gobernanza moderna. A diferencia de los entornos lúdicos o simulados donde el RL tradicional ha logrado éxitos históricos mediante la exploración irrestricta, el ámbito de la administración pública exige garantías estrictas de seguridad, interpretabilidad y cumplimiento de presupuestos finitos. En este contexto, los Procesos de Decisión de Markov Restringidos, conocidos en la literatura como “Constrained Markov Decision

Processes” (CMDPs), emergen como el marco matemático natural para formular políticas públicas. En un CMDP estándar, el objetivo es maximizar una función de recompensa objetivo, sujeta a restricciones de costo. Formalmente, Feinberg, Jaśkiewicz y Nowak (2019) establecen un modelo donde se busca $\max_{\pi} J_0(\pi)$ sujeto a $J_i(\pi) \geq d_i$ para $i \in I$, donde $J_i(\pi) = \mathbb{E}_{\mu}^{\pi}[\sum_{n=1}^{\infty} \beta^{n-1} r_i(s_n, a_n)]$ representa el valor esperado descontado de las recompensas o costos r_i bajo una política π y un factor de descuento β . Al demostrar que estos modelos admiten políticas óptimas estacionarias y deterministas en espacios de estados de Borel con funciones de recompensa no acotadas, proveen la garantía teórica de que un gobierno puede, matemáticamente, encontrar una estrategia óptima y estable en entornos continuos (como la asignación monetaria de subsidios a lo largo del tiempo) sin violar las restricciones legales impuestas (d_i). Complementando esta visión orientada al descuento, Yu (2021) aborda los CMDPs desde la perspectiva del costo promedio a largo plazo mediante programas lineales de dimensión infinita. Su modelo plantea el problema dual de optimización $\inf_{\gamma} \langle \gamma, c_0 \rangle$ sujeto a $L\gamma = b$, donde γ representa la medida de ocupación (la frecuencia con la que un sistema visita un par estado-acción a largo plazo). Al probar la ausencia de brechas de dualidad en estos modelos bajo costos estrictamente no acotados, se habilita el uso de métodos de programación lineal para encontrar pares óptimos de estado-política. Para la gestión pública, esto es vital en el diseño de infraestructuras sostenibles o redes de transporte, donde el rendimiento promedio y el respeto continuo a restricciones de capacidad a lo largo de décadas son más relevantes que las recompensas descontadas a corto plazo.

Para que estos modelos teóricos se traduzcan en herramientas viables para los tomadores de decisiones gubernamentales, la eficiencia muestral es un requisito ineludible, dado que interactuar con la sociedad para “recolectar datos” es sumamente costoso y éticamente delicado. Kalagarla, Jain y Nuzzo (2021) formulan un modelo para CMDPs episódicos de horizonte finito que logra garantías de ser “Probablemente Aproximadamente Correcto” (PAC). Su algoritmo en línea requiere una complejidad muestral delimitada por $\tilde{O}(|S||A|C^2H^2/\epsilon^2)$, donde $|S|$ y $|A|$ son los tamaños del espacio de estados y acciones, H es el horizonte temporal, y ϵ es el margen de error tolerado. Esta formulación asegura que un programa piloto del gobierno (por ejemplo, una intervención educativa de duración H) pueda aprender una política casi óptima sin requerir una experimentación poblacional infinita. En una línea paralela enfocada en múltiples restricciones simultáneas, Buckley, Papathanasiou y Spanopoulos (2025) derivan límites de complejidad muestral utilizando un enfoque primal-dual basado

en modelos generativos. Su avance matemático radica en modelar explícitamente la dicotomía entre viabilidad relajada y estricta. En su modelo, la viabilidad estricta requiere que el valor del costo de la política cumpla inquebrantablemente $\hat{V}_{c_i}^{\pi}(\rho) \geq b_i$, lo que matemáticamente exige una complejidad muestral mucho mayor, de orden $\tilde{O}(d^3|S||A|/(\epsilon^2\zeta_c^{*2}))$, donde ζ_c^* es la brecha de Slater que mide el margen de holgura de la restricción. Para los formuladores de políticas, esta distinción permite calibrar el riesgo: un programa de optimización de tráfico autónomo requeriría el marco de viabilidad estricta para garantizar cero colisiones, mientras que un sistema de asignación de becas podría operar bajo viabilidad relajada $\hat{V}_{c_i}^{\pi}(\rho) \geq b_i - \epsilon$ para acelerar su despliegue a costa de excederse marginalmente en el presupuesto inicial.

Sin embargo, el mundo real en el que operan las políticas públicas rara vez es estático; está caracterizado por shocks económicos, cambios demográficos y crisis imprevistas que alteran las dinámicas subyacentes. Stradi et al. (2024) abordan esta no estacionariedad desarrollando algoritmos para CMDPs con recompensas y restricciones que varían de forma adversaria. Introducen matemáticamente un parámetro de corrupción $C = \max\{C_G, C_r\}$ que cuantifica la distancia máxima entre las distribuciones de recompensas reales y unas teóricas estacionarias a lo largo de los episodios. Al lograr un arrepentimiento delimitado por $\tilde{O}(\sqrt{T} + C)$, demuestran que un sistema de IA puede degradar su rendimiento de manera elegante y controlada cuando las condiciones sociales cambian abruptamente, una característica indispensable para sistemas de triaje hospitalario durante una pandemia, donde los costos y capacidades fluctúan sin previo aviso. Elevando el tratamiento de la incertidumbre, Zuo y He (2025) introducen el algoritmo “SPOT” (“Stochastic Pessimistic-Optimistic Thresholding”), donde el límite de seguridad mismo se modela como un proceso estocástico. En su formulación, el agente no observa un umbral fijo α_i , sino señales ruidosas $\tilde{\alpha}_{i,h}(s, a)$, siendo el umbral latente la suma esperada $\alpha_i = \sum_{h=1}^H \mathbb{E}[\tilde{\alpha}_{i,h}]$. Esto es crítico para el diseño de políticas ambientales o de salud pública donde el límite exacto de toxicidad o la tolerancia fiscal de los ciudadanos no se conoce a priori y debe ser inferido dinámicamente, garantizando matemáticamente que el algoritmo respete los límites dinámicos de la sociedad conforme estos se vuelven evidentes.

A medida que los espacios de estados de la administración pública se vuelven masivos, los enfoques tabulares dejan de ser prácticos, haciendo imperativo el uso de aproximación de funciones mediante arquitecturas de aprendizaje profundo. Xu et al. (2025) proponen un algoritmo de “Actor-Crítico Natural Primal-Dual” para CMDPs

de recompensa promedio utilizando parametrización general θ . Su modelo optimiza una función Lagrangiana $\mathcal{L}(\theta, \lambda) = J_r(\theta) + \lambda J_c(\theta)$, donde las actualizaciones de la política siguen el gradiente de política natural $\omega_{g,\theta}^* = F(\theta)^{-1} \nabla_{\theta} J_g(\theta)$, siendo $F(\theta)$ la matriz de información de Fisher. Al demostrar matemáticamente una convergencia global de $\tilde{O}(1/\sqrt{T})$, permiten que el sector público utilice redes neuronales complejas para procesar datos de alta dimensionalidad (como imágenes satelitales o registros censales) con la garantía de que el modelo convergerá a una política segura y óptima. Para hacer este despliegue más seguro y rápido, el paradigma de aprendizaje “Offline-to-Online” (O2O) se vuelve crucial, permitiendo a la IA aprender de datos históricos antes de interactuar con el mundo real. Chen et al. (2026) identifican que el ajuste fino ingenuo (“fine-tuning”) de políticas falla en CMDPs debido al desajuste del multiplicador de Lagrange y a las estimaciones erróneas de los valores Q. Su marco “Marvel” resuelve esto mediante una etapa de prealineación de valor (VPA) que minimiza la pérdida $\mathcal{L}_Q^{\text{VPA}}(\omega) = \mathbb{E}_{\mathcal{D}}[(Q(s,a) - (r + \gamma \mathbb{E}[\hat{Q}(s',a') - \alpha \log \pi_0(a'|s')]))^2]$. Matemáticamente, este término de penalización entrópica $\alpha \log \pi_0$ obliga a las estimaciones a corregirse utilizando solo datos offline, previniendo que la IA asigne costos falsamente altos a acciones fuera de distribución que podrían ser beneficiosas. En la práctica gubernamental, esto permite inicializar un modelo de IA con datos de registros administrativos de la última década y desplegarlo en vivo de forma segura, superando la inercia algorítmica conservadora que impediría mejorar el programa social.

Dado que las políticas públicas deben basarse mayoritariamente en datos observacionales debido a la inviabilidad ética de la experimentación aleatorizada a gran escala, la extracción causal rigurosa de estos datos es imperativa. Athey y Wager (2021) desarrollan un método matemático para el aprendizaje de políticas a partir de datos observacionales utilizando estimación doblemente robusta. Su estimador para el valor de una política se define como $\hat{V}(\pi) = \frac{1}{n} \sum_{i=1}^n \hat{\Gamma}_i$, donde el puntaje corregido es $\hat{\Gamma}_i = \hat{\mu}(X_i, 1) + \frac{W_i}{\hat{e}(X_i)} (Y_i - \hat{\mu}(X_i, 1))$ (para el grupo tratado, y análogamente para control). Este modelo neutraliza matemáticamente los sesgos de selección (endogeneidad) presentes en los datos históricos, garantizando que los formuladores de políticas puedan evaluar el verdadero impacto causal de asignar subsidios o intervenciones policiales. Extendiéndolo a dimensiones dinámicas, Ertefaie y Strawderman (2018) utilizan ecuaciones de Bellman con aprendizaje Q para optimizar regímenes de tratamiento sobre horizontes de tiempo indefinidos, donde el error de diferencia temporal se define como $\delta_{t+1}(\theta) = R_{t+1} - Q_{\theta}(S_t, A_t) + \gamma \max_{a'} Q_{\theta}(S_{t+1}, a')$. Este enfoque de

inducción hacia atrás es fundamental para la gestión de enfermedades crónicas o programas de bienestar social prolongados, asegurando que las decisiones de hoy optimicen matemáticamente las trayectorias de salud y económicas de los ciudadanos a lo largo de toda su vida, sin extrapolar irracionalmente más allá del horizonte temporal observable.

Para que los algoritmos derivados de estos datos se desplieguen sin riesgo, la evaluación fuera de política (“Off-Policy Evaluation”, OPE) debe venir con intervalos de confianza rigurosos. Thomas, Theodorou y Ghavamzadeh (2015) formulan estadísticamente la evaluación fuera de política de alta confianza, buscando un límite inferior $\hat{V}_{\text{lower}}$ tal que $\Pr(V(\pi_e) \geq \hat{V}_{\text{lower}}) \geq 1 - \delta$. Utilizando desigualdades de concentración empíricas de Bernstein aplicadas a pesos de muestreo de importancia, su modelo actúa como una red de seguridad matemática, el equivalente a un estudio de impacto ambiental, asegurando a los legisladores con un 95 % de certeza que una nueva intervención algorítmica no será perjudicial. Incluso durante la propia exploración, la seguridad debe estar garantizada. Wendl et al. (2026) desarrollan “SOOPER”, que utiliza modelos dinámicos probabilísticos y un límite de costo pesimista definido como $\bar{V}_c^{\hat{\pi}}(s) = \max_{\tilde{f} \in \mathcal{F}_n} \mathbb{E}_{\hat{\pi}}[\sum \gamma^t c(s_t, a_t)]$, donde $\mathcal{F}_n$ es un conjunto de modelos plausibles. Si la IA estima que la exploración activa violará el presupuesto bajo el peor modelo dinámico posible, el sistema revierte matemáticamente al “prior” conservador. En la infraestructura urbana, esto significa que un algoritmo de semaforización intentará flujos innovadores, pero si el modelo matemático detecta un riesgo inminente de saturación en el peor escenario probabilístico, revertirá inmediatamente al patrón clásico probado, garantizando fluidez sin riesgos.

Más allá del aprendizaje por refuerzo tradicional, las políticas públicas deben diferenciar claramente entre correlación y causalidad. Xiao et al. (2026) logran esto integrando la inferencia causal en políticas de difusión (“Diffusion Models”) mediante su marco “CausalGDP”. Modifican la función de puntuación de la difusión añadiendo un término de orientación causal: $\nabla_a \log p(s_{t+1}, r_t \mid s_t, \text{do}(a_t))$. La inserción del operador $\text{do}(\cdot)$ de Pearl asegura matemáticamente que la IA genere acciones basándose únicamente en el impacto causal directo sobre las recompensas futuras, bloqueando las variables de confusión (como el estatus socioeconómico en la predicción de criminalidad). De forma complementaria, De Moor et al. (2026) presentan “SCOPE”, que modela la inferencia causal secuencial definiendo la función Q como $Q_k(\sigma_{:l_k}, a_k) = \mathbb{E}[V_{k+1}(\sigma_{:l_{k+1}}) \mid \sigma_{:l_k}, a_k]$. Este modelo de inducción hacia atrás causal

permite al gobierno predecir con precisión matemática cómo una secuencia de intervenciones, por ejemplo, enviar un recordatorio tributario seguido de una auditoría aleatoria, causará de manera secuencial y dependiente un aumento en la recaudación fiscal, optimizando recursos públicos escasos con intervenciones quirúrgicas y demostrables.

El diseño de IA para el sector público requiere intrínsecamente la humildad algorítmica: la capacidad matemática de abstenerse y delegar a un humano cuando la incertidumbre es alta. Sawarni et al. (2026) formalizan esto en su modelo de “Aprendizaje de Políticas con Abstención”. Modifican la función de valor para incluir un tercer estado de acción (*), definiéndola como $V^{(p)}(\pi) = \mathbb{E}[\mathbf{1}\{\pi(X) \neq *\}v(\pi, X) + \mathbf{1}\{\pi(X) = *\}g_o(*, X)]$, donde $g_o(*, X) = \mathbb{E}[\frac{Y(1)+Y(0)}{2} + p \mid X]$. En este modelo, el algoritmo recibe el valor de una suposición aleatoria más un bono explícito p cuando elige abstenerse (*). Esta penalización y recompensa estructural asegura matemáticamente que, en regiones del espacio de características donde el margen causal es estrecho o la varianza es extrema, la política óptima calculada sea abstenerse. En la práctica gubernamental, esto permite que un sistema automatizado de procesamiento de asilo o asignación de planes de asistencia social apruebe o rechace automáticamente los casos estadísticamente claros y seguros, pero delegue matemáticamente a jueces humanos o trabajadores sociales los casos ambiguos o atípicos. Esta formulación no solo proporciona garantías de arrepentimiento de orden $\mathcal{O}(1/n)$, sino que funda una gobernanza pública aumentada por IA donde la robustez distribucional, la equidad, el cumplimiento presupuestario estricto y la rendición de cuentas humana coexisten dentro de un mismo tejido algorítmico y matemático, garantizando políticas públicas eficientes y socialmente responsables.

3. Marco formal, notación y arquitectura matemática de la IA pública

Una IA para optimización de políticas públicas no se reduce a un predictor de resultados ni a un optimizador sin restricciones. Debe ser modelada como una composición de módulos con supuestos distinguibles, errores cuantificables y parámetros normativos auditables. La formulación adoptada en este documento no pretende ser una teoría completa de gobierno algorítmico, sino una formalización rigurosa de una clase de sistemas de apoyo a decisión pública donde la decisión se expresa como una

política dinámica factible y donde la inferencia causal, la robustez y la seguridad son partes constitutivas del problema y no añadidos ex post.

El núcleo analítico se apoya en procesos de decisión de Markov (MDP) finitos con descuento, representación por medidas ocupacionales y programación lineal restringida. Esta elección proporciona un lenguaje suficientemente general para incorporar dinámica, incertidumbre y restricciones de política pública. Sin embargo, el documento explicita desde el inicio que el MDP finito estacionario es una aproximación del sistema real, cuya validez depende de supuestos de agregación, observabilidad y estabilidad estructural. La teoría desarrollada permite demostrar resultados exactos en el modelo y, además, incorporar cotas que cuantifican el error cuando se trabaja con un modelo estimado o robustificado.

Se introduce una arquitectura de tres capas. La primera es la capa causal-estadística, que produce estimaciones de transiciones, recompensas y restricciones a partir de datos observacionales o experimentales, junto con descripciones de incertidumbre. La segunda es la capa de optimización, que resuelve un problema dinámico restringido sobre el modelo estimado, incluyendo dualidad y precios sombra. La tercera es la capa normativa-institucional, donde se fijan pesos sociales, umbrales de equidad, presupuestos, reglas de seguridad y límites de implementabilidad. La separación es sustantiva, porque una mejora exacta en la segunda capa puede ser irrelevante o dañina si la primera capa está sesgada o si la tercera capa está mal especificada.

Definición 1. Una *IA pública restringida* se modela como un operador

$$\mathcal{A}_n : (\mathcal{D}_n, \mathcal{C}) \mapsto \hat{\pi}_n,$$

donde $\mathcal{D}_n$ es la información disponible (datos administrativos, encuestas, registros, experimentos, series temporales, etc.), $\mathcal{C}$ es el conjunto de parámetros normativos e institucionales (pesos, umbrales, restricciones, política base, reglas de seguridad), y $\hat{\pi}_n$ es una política de decisión factible.

En esta definición, la IA no se identifica con una arquitectura computacional específica (árboles, modelos lineales, redes neuronales, métodos bayesianos o híbridos), sino con un operador de decisión con estructura modular. Esta abstracción permite derivar resultados matemáticos que dependen de propiedades del problema (convexidad, soporte, contracción, regularidad) y no de una implementación particular.

Definición 2. Un *MDP de política pública* es una quintupla

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma),$$

donde $\mathcal{S}$ es un conjunto finito de estados, $\mathcal{A}$ un conjunto finito de acciones, $P(s'|s, a)$ una probabilidad de transición, $r(s, a) \in \mathbb{R}$ una recompensa social inmediata y $\gamma \in (0, 1)$ un factor de descuento.

En aplicaciones públicas, s puede resumir condiciones macroeconómicas, sectoriales, distributivas, institucionales y ambientales; a puede representar intensidad de transferencias, asignaciones de inversión, parámetros regulatorios, esquemas de fiscalización, reglas de focalización u otras combinaciones de instrumentos discretizados. La finitud de $\mathcal{S}$ y $\mathcal{A}$ responde a un propósito de tractabilidad y auditabilidad, no a una afirmación ontológica sobre el sistema real.

Definición 3. Una política estocástica estacionaria es una familia $\pi(a|s)$ tal que $\pi(a|s) \geq 0$ y $\sum_{a \in \mathcal{A}} \pi(a|s) = 1$ para todo $s \in \mathcal{S}$. El conjunto de todas ellas se denota por Π .

Se fija una distribución inicial μ sobre $\mathcal{S}$. El valor social descontado de una política π es

$$J(\pi) = \mathbb{E}_{s_0 \sim \mu, \pi, P} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right].$$

Una escalarización operativa de objetivos múltiples puede escribirse como

$$r(s, a) = \sum_{j=1}^q \alpha_j \tilde{f}_j(s, a), \quad \alpha_j \geq 0,$$

donde $\tilde{f}_j$ son componentes normalizados o reescalados de resultados (actividad, pobreza, desigualdad, cobertura, externalidades, costo fiscal, etc.) y α_j son pesos normativos auditables.

La normalización es central para la interpretabilidad numérica de α_j . Si las magnitudes f_j están en unidades heterogéneas, el uso de pesos directamente sobre f_j produce parámetros cuya escala depende de la unidad de medición. Para evitar esa

ambigüedad, se introducen transformaciones como

$$\tilde{f}_j(s, a) = \frac{f_j(s, a) - m_j}{\sigma_j} \quad \text{o} \quad \tilde{f}_j(s, a) \in [0, 1],$$

con m_j, σ_j elegidos por reglas transparentes. La matemática no fija la norma de normalización, pero exige que sea explícita si se pretende comparar o auditar pesos.

La arquitectura modular se formaliza mediante la composición

$$\mathcal{D}_n \xrightarrow{\Psi_n} (\hat{P}, \hat{r}, \hat{g}, \hat{\mathcal{U}}, \hat{\mathcal{I}}) \xrightarrow{\mathcal{O}} \hat{\pi} \xrightarrow{\mathcal{S}} \pi^{\text{dep}},$$

donde Ψ_n es el módulo causal-estadístico, $\mathcal{O}$ es el optimizador restringido y $\mathcal{S}$ representa el filtro de seguridad/despliegue que puede rechazar o regularizar la política candidata. El objeto $\hat{\mathcal{U}}$ codifica incertidumbre de modelo (por ejemplo conjuntos robustos) y $\hat{\mathcal{I}}$ contiene elementos de inferencia para seguridad estadística (cotas, intervalos, bandas).

La diferencia entre optimalidad en el modelo estimado y desempeño en el sistema real se descompone de forma útil. Si π^* es óptima para el modelo verdadero M y $\hat{\pi}$ es la política producida por la IA sobre un modelo estimado $\hat{M}$, entonces

$$J_M(\pi^*) - J_M(\hat{\pi}) = \underbrace{J_M(\pi^*) - J_{\hat{M}}(\pi^*)}_{\text{error de modelo en } \pi^*} + \underbrace{J_{\hat{M}}(\pi^*) - J_{\hat{M}}(\hat{\pi})}_{\text{error de optimización}} + \underbrace{J_{\hat{M}}(\hat{\pi}) - J_M(\hat{\pi})}_{\text{error de transferencia}}.$$

Esta identidad es algebraicamente trivial pero conceptualmente decisiva: evita atribuir a la capa de optimización problemas que pertenecen a la capa causal-estadística o al cambio de régimen entre entorno estimado y entorno real.

Observación 1. Cuando el sistema real es parcialmente observable, una historia H_t induce un estado de creencia b_t sobre estados latentes y el problema natural es un POMDP. El MDP sobre estados agregados $S_t = \phi(H_t)$ debe entenderse como proyección aproximada. Las garantías exactas de este documento se refieren al MDP modelado, mientras que las garantías respecto al sistema real requieren hipótesis adicionales de agregación suficiente.

Se adopta una política de notación explícita para reducir ambigüedad entre módulos. Las variables de historia causal se denotarán por H_t, X_t, A_t , las variables del MDP agregado por S_t, A_t cuando la acción coincida, los funcionales de objetivos por $F_j(\pi)$,

los resultados grupales por $J_g(\pi)$ o $\mathbb{E}[Y_g^\pi]$, y las restricciones por $G_k(\pi) \leq b_k$. La distancia de variación total se define con la convención estándar de factor $1/2$:

$$\text{TV}(p, q) = \frac{1}{2} \sum_x |p(x) - q(x)|.$$

Símbolo	Tipo	Interpretación
$\mathcal{D}_n$	Datos	Base empírica disponible para estimación e inferencia
$\mathcal{C}$	Parámetros	Restricciones, pesos normativos, política base, reglas de seguridad
$\hat{\Psi}_n$	Operador	Módulo causal-estadístico que produce $\hat{P}, \hat{r}, \hat{g}$ y objetos de incertidumbre
$\mathcal{O}$	Operador	Optimizador restringido sobre el modelo estimado/robusto
$\hat{\pi}$	Política	Política candidata producida por la IA
π_b	Política	Política base o régimen vigente
$P, \hat{P}$	Transición	Modelo verdadero / estimado de dinámica bajo intervención
$r, \hat{r}$	Recompensa	Función social inmediata verdadera / estimada
$G_k(\pi)$	Funcional	Restricción pública agregada (presupuesto, equidad, capacidad, etc.)
$\mathcal{P}_{s,a}$	Conjunto	Incertidumbre robusta de transiciones en (s, a)
TV	Distancia	Variación total, $\frac{1}{2} \ p - q\ _1$
KL	Divergencia	Divergencia relativa para estabilidad y regularización

Fuente: Elaboración propia.

4. Formulación dinámica restringida, medidas ocupacionales, realizabilidad, dualidad y mejora de política

La representación por medidas ocupacionales descontadas convierte el problema dinámico en un programa lineal finito cuando $\mathcal{S}$ y $\mathcal{A}$ son finitos. Esta transformación es particularmente útil en política pública porque permite incorporar restricciones de presupuesto, equidad, implementabilidad y estabilidad institucional de forma lineal o linealizables, sin perder interpretabilidad de los multiplicadores duales.

Definición 4. Dada una política $\pi \in \Pi$, su *medida ocupacional descontada* es

$$d_\pi(s, a) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_\pi(S_t = s, A_t = a).$$

Proposición 1. Para toda $\pi \in \Pi$, la medida d_π satisface

$$d_\pi(s, a) \geq 0, \quad \sum_{s, a} d_\pi(s, a) = 1,$$

y las ecuaciones de flujo

$$\sum_a d_\pi(s, a) = (1 - \gamma)\mu(s) + \gamma \sum_{s', a'} d_\pi(s', a') P(s|s', a').$$

Demostración. La no negatividad es inmediata por construcción. Para la suma total,

$$\sum_{s, a} d_\pi(s, a) = (1 - \gamma) \sum_{t \geq 0} \gamma^t \sum_{s, a} \mathbb{P}_\pi(S_t = s, A_t = a) = (1 - \gamma) \sum_{t \geq 0} \gamma^t = 1.$$

Para el flujo, se tiene

$$\sum_a d_\pi(s, a) = (1 - \gamma) \sum_{t \geq 0} \gamma^t \mathbb{P}_\pi(S_t = s).$$

Separando el término $t = 0$,

$$\sum_a d_\pi(s, a) = (1 - \gamma)\mu(s) + (1 - \gamma) \sum_{t \geq 1} \gamma^t \mathbb{P}_\pi(S_t = s).$$

Reindexando con $u = t - 1$,

$$= (1 - \gamma)\mu(s) + \gamma(1 - \gamma) \sum_{u \geq 0} \gamma^u \mathbb{P}_\pi(S_{u+1} = s).$$

Por la ley total de probabilidad y la dinámica del MDP,

$$\mathbb{P}_\pi(S_{u+1} = s) = \sum_{s', a'} \mathbb{P}_\pi(S_u = s', A_u = a') P(s|s', a').$$

Sustituyendo y permutando sumas,

$$\sum_a d_\pi(s, a) = (1 - \gamma)\mu(s) + \gamma \sum_{s', a'} d_\pi(s', a') P(s|s', a').$$

□

Proposición 2. *El valor de π se expresa linealmente en d_π :*

$$J(\pi) = \frac{1}{1 - \gamma} \sum_{s, a} d_\pi(s, a) r(s, a).$$

Demostración. Por definición de d_π ,

$$\sum_{s, a} d_\pi(s, a) r(s, a) = (1 - \gamma) \sum_{t \geq 0} \gamma^t \mathbb{E}_\pi[r(S_t, A_t)] = (1 - \gamma) J(\pi).$$

Dividiendo por $1 - \gamma$, se obtiene la identidad.

□

Las restricciones públicas se modelan mediante funcionales lineales o convexos sobre d_π . Una forma general para restricciones lineales es

$$G_k(\pi) = \frac{1}{1 - \gamma} \sum_{s, a} d_\pi(s, a) g_k(s, a) \leq b_k, \quad k = 1, \dots, m.$$

Los términos g_k pueden representar costo fiscal, carga administrativa, riesgo ambiental, consumo de capacidad, intensidad de cambio respecto a un régimen base o magnitudes distributivas linearizadas.

La formulación primal en medidas ocupacionales queda

$$\max_d \quad \frac{1}{1-\gamma} \sum_{s,a} d(s,a) r(s,a)$$

sujeta a

$$\sum_a d(s,a) = (1-\gamma)\mu(s) + \gamma \sum_{s',a'} d(s',a') P(s|s',a') \quad \forall s \in \mathcal{S},$$

$$d(s,a) \geq 0 \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A},$$

$$\frac{1}{1-\gamma} \sum_{s,a} d(s,a) g_k(s,a) \leq b_k \quad \forall k = 1, \dots, m.$$

Teorema 1. Si $\mathcal{S}$ y $\mathcal{A}$ son finitos, r y g_k son acotadas y el conjunto factible es no vacío, entonces el problema primal admite una solución óptima.

Demostración. Las restricciones definen una intersección de semiespacios y un subespacio afín en $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. Además, toda solución factible satisface $\sum_{s,a} d(s,a) = 1$, por lo que el conjunto factible está contenido en el simplex y es acotado. Como es cerrado, resulta compacto. El objetivo es lineal y por tanto continuo; la existencia del máximo se sigue por Weierstrass. $\square$

La realizabilidad de medidas ocupacionales factibles es crucial para cerrar el puente entre el programa lineal y el conjunto de políticas estacionarias.

Proposición 3 (Realizabilidad de medidas ocupacionales). Sea d una medida que satisfice no negatividad, suma unitaria y ecuaciones de flujo. Defínase

$$\pi_d(a|s) = \begin{cases} \frac{d(s,a)}{\sum_{a'} d(s,a')}, & \text{si } \sum_{a'} d(s,a') > 0, \\ \bar{\pi}(\cdot|s), & \text{si } \sum_{a'} d(s,a') = 0, \end{cases}$$

donde $\bar{\pi}(\cdot|s)$ es una distribución arbitraria. Entonces d es la medida ocupacional descontada de π_d . En particular, no se pierde optimalidad al restringirse a políticas estacionarias aleatorias en el MDP finito descontado.

Demostración. Sea $x(s) = \sum_a d(s,a)$. Por definición de π_d , se tiene

$$d(s,a) = x(s)\pi_d(a|s) \quad \text{cuando } x(s) > 0.$$

Si $x(s) = 0$, entonces $d(s, a) = 0$ para todo a . Las ecuaciones de flujo implican

$$x(s) = (1 - \gamma)\mu(s) + \gamma \sum_{s'} x(s') \sum_{a'} \pi_d(a'|s') P(s|s', a').$$

Esta ecuación coincide con la ecuación de balance descontado de las probabilidades marginales ocupacionales inducidas por π_d . Por unicidad de la solución del sistema lineal de balance descontado, las marginales ocupacionales de π_d son $x(s)$, y por la factorización $d(s, a) = x(s)\pi_d(a|s)$, se concluye que $d = d_{\pi_d}$. $\square$

Esta proposición cierra una objeción frecuente: la programación lineal sobre medidas ocupacionales no introduce objetos espurios en el caso finito descontado, sino una representación equivalente del problema dinámico en la clase de políticas estacionarias aleatorias.

En forma matricial, sea $n = |\mathcal{S}||\mathcal{A}|$, sea $d \in \mathbb{R}^n$ el vector de medidas, sea $c \in \mathbb{R}^n$ con componentes $c(s, a) = \frac{1}{1-\gamma}r(s, a)$, sea $A \in \mathbb{R}^{|\mathcal{S}| \times n}$ la matriz de flujo y $G \in \mathbb{R}^{m \times n}$ la matriz de restricciones con

$$(Ad)(s) = \sum_a d(s, a) - \gamma \sum_{s', a'} d(s', a') P(s|s', a'), \quad (Gd)_k = \frac{1}{1-\gamma} \sum_{s, a} d(s, a) g_k(s, a).$$

Entonces el primal se escribe

$$\max_d c^\top d \quad \text{s.a.} \quad Ad = (1 - \gamma)\mu, \quad Gd \leq b, \quad d \geq 0.$$

El dual asociado introduce un multiplicador $v \in \mathbb{R}^{|\mathcal{S}|}$ para la restricción de flujo y $\lambda \in \mathbb{R}_+^m$ para las desigualdades:

$$\min_{v, \lambda \geq 0} (1 - \gamma)\mu^\top v + b^\top \lambda$$

sujeto a

$$A^\top v + G^\top \lambda \geq c.$$

Desplegando la restricción coordinada por coordinada se obtiene, para todo (s, a) ,

$$v(s) - \gamma \sum_{s'} P(s'|s, a) v(s') + \frac{1}{1-\gamma} \sum_{k=1}^m \lambda_k g_k(s, a) \geq \frac{1}{1-\gamma} r(s, a),$$

equivalentemente,

$$v(s) \geq \frac{1}{1-\gamma} \left(r(s, a) - \sum_k \lambda_k g_k(s, a) \right) + \gamma \sum_{s'} P(s'|s, a) v(s').$$

Definiendo la recompensa penalizada $r_\lambda(s, a) = r(s, a) - \sum_k \lambda_k g_k(s, a)$, la restricción dual adopta la forma de una desigualdad de Bellman penalizada:

$$v(s) \geq \frac{1}{1-\gamma} r_\lambda(s, a) + \gamma \sum_{s'} P(s'|s, a) v(s').$$

Teorema 2 (Dualidad fuerte). *Si el primal es factible, entonces el valor primal coincide con el valor dual. Además, si el valor óptimo es finito, existen soluciones óptimas primal y dual.*

Demostración. El problema primal y su dual son programas lineales en dimensión finita. El conjunto factible primal es no vacío y acotado por estar contenido en el simplex. El objetivo es finito por acotación de r . Por el teorema de dualidad fuerte de programación lineal, la brecha dual es cero y existen soluciones óptimas en ambos problemas. $\square$

Proposición 4 (Condiciones KKT en forma primal-dual). *Sea d^* solución óptima primal y (v^*, λ^*) solución óptima dual. Entonces se satisfacen: factibilidad primal, factibilidad dual, complementariedad dual-primal en restricciones*

$$\lambda_k^* ((Gd^*)_k - b_k) = 0, \quad k = 1, \dots, m,$$

y complementariedad por coordenada

$$d^*(s, a) \left(v^*(s) - \gamma \sum_{s'} P(s'|s, a) v^*(s') - \frac{1}{1-\gamma} r_{\lambda^*}(s, a) \right) = 0.$$

Demostración. Las dos primeras partes corresponden a factibilidad. La complementariedad $\lambda_k^* ((Gd^*)_k - b_k) = 0$ es la condición estándar de LP entre multiplicadores no negativos y desigualdades. La complementariedad por coordenada proviene de la condición de holgura complementaria entre $d \geq 0$ y la restricción dual $A^\top v + G^\top \lambda \geq c$, expresada en coordenadas (s, a) . $\square$

La interpretación institucional es directa: λ_k^* es el precio sombra de la restricción k . Si

una restricción no es vinculante, su precio sombra es nulo. Si una restricción presiona el óptimo, su multiplicador mide el costo marginal de endurecerla. Esta interpretación es una de las razones por las que la formulación primal-dual es adecuada para auditoría de decisiones públicas.

Se introducen ahora los objetos de valor y ventaja para resultados de mejora de política. Para $\pi \in \Pi$,

$$V^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t) \mid S_0 = s \right], \quad Q^\pi(s, a) = r(s, a) + \gamma \sum_{s'} P(s'|s, a) V^\pi(s').$$

El operador de Bellman de política es

$$(T^\pi V)(s) = \sum_a \pi(a|s) \left(r(s, a) + \gamma \sum_{s'} P(s'|s, a) V(s') \right).$$

Lema 1. *El operador T^π es una contracción en norma supremo con módulo γ :*

$$\|T^\pi V - T^\pi W\|_\infty \leq \gamma \|V - W\|_\infty.$$

Demostración. Para cada s ,

$$|(T^\pi V)(s) - (T^\pi W)(s)| \leq \sum_a \pi(a|s) \gamma \sum_{s'} P(s'|s, a) |V(s') - W(s')| \leq \gamma \|V - W\|_\infty.$$

Tomando supremo en s , se concluye. □

Corolario 1. *Para toda política π , existe un único punto fijo V^π de T^π .*

Teorema 3 (Mejora de política). *Si π' satisface*

$$\sum_a \pi'(a|s) Q^\pi(s, a) \geq V^\pi(s) \quad \forall s \in \mathcal{S},$$

entonces $V^{\pi'}(s) \geq V^\pi(s)$ para todo s , y en particular $J(\pi') \geq J(\pi)$.

Demostración. La hipótesis equivale a $(T^{\pi'} V^\pi)(s) \geq V^\pi(s)$ para todo s . Por monotonidad de $T^{\pi'}$,

$$V^\pi \leq T^{\pi'} V^\pi \leq (T^{\pi'})^2 V^\pi \leq \dots$$

La contracción de $T^{\pi'}$ implica $(T^{\pi'})^n V^{\pi} \rightarrow V^{\pi'}$ en norma supremo. Pasando al límite, se obtiene $V^{\pi} \leq V^{\pi'}$ coordenada a coordenada. $\square$

Se define la ventaja de π como

$$A^{\pi}(s, a) = Q^{\pi}(s, a) - V^{\pi}(s).$$

El siguiente resultado es central para seguridad y mejora conservadora.

Teorema 4 (Lema de diferencia de desempeño). *Para cualesquiera $\pi, \pi' \in \Pi$,*

$$J(\pi') - J(\pi) = \frac{1}{1 - \gamma} \mathbb{E}_{(s,a) \sim d_{\pi'}} [A^{\pi}(s, a)].$$

Demostración. Partiendo de la definición de ventaja,

$$A^{\pi}(s, a) = r(s, a) + \gamma \sum_{s'} P(s'|s, a) V^{\pi}(s') - V^{\pi}(s).$$

Tomando esperanza con respecto a $d_{\pi'}$ y multiplicando por $1/(1 - \gamma)$,

$$\begin{aligned} \frac{1}{1 - \gamma} \mathbb{E}_{d_{\pi'}} [A^{\pi}] &= \frac{1}{1 - \gamma} \sum_{s,a} d_{\pi'}(s, a) r(s, a) \\ &\quad + \frac{\gamma}{1 - \gamma} \sum_{s,a} d_{\pi'}(s, a) \\ &\quad \quad \times \sum_{s'} P(s'|s, a) V^{\pi}(s') \\ &\quad - \frac{1}{1 - \gamma} \sum_s \left(\sum_a d_{\pi'}(s, a) \right) V^{\pi}(s). \end{aligned}$$

El primer término es $J(\pi')$. Para el último, se usa la ecuación de flujo de $d_{\pi'}$:

$$\sum_a d_{\pi'}(s, a) = (1 - \gamma) \mu(s) + \gamma \sum_{s', a'} d_{\pi'}(s', a') P(s|s', a').$$

Sustituyendo, los términos con $\gamma P V^{\pi}$ se cancelan y queda

$$\frac{1}{1 - \gamma} \mathbb{E}_{d_{\pi'}} [A^{\pi}] = J(\pi') - \sum_s \mu(s) V^{\pi}(s) = J(\pi') - J(\pi).$$

$\square$

Corolario 2. Si $\mathbb{E}_{d_{\pi'}}[A^{\pi}] \geq 0$, entonces $J(\pi') \geq J(\pi)$.

Este lema se usará más adelante para formular seguridad respecto a una política base y para vincular mejora de política con estimación de ventajas bajo incertidumbre.

Objeto	Forma matemática	Función en el marco
Medida ocupacio- nal	$d_{\pi}(s, a)$	Representa dinámica descontada como LP finito
Restricción pública	$G_k(\pi) \leq b_k$	Incorpora presupuesto, capacidad, equidad, riesgo
Dualidad	(v, λ)	Explica precios sombra y Bellman penalizado
Realizabilidad	$d \mapsto \pi_d$	Garantiza equivalencia LP- política en MDP finito
Ventaja	$A^{\pi} = Q^{\pi} - V^{\pi}$	Mide ganancia relativa de accio- nes frente a π
Lema de desempe- ño	$J(\pi') - J(\pi) = \frac{1}{1-\gamma} \mathbb{E}_{d_{\pi'}}[A^{\pi}]$	Base de seguridad y mejora con- servadora

Fuente: Elaboración propia.

5. Equidad, implementabilidad, estabilidad institucional, multiob- jetivo y sensibilidad normativa

La estructura de restricciones públicas no se agota en presupuesto. En políticas públicas, una IA puede ser formalmente óptima y, al mismo tiempo, políticamente inaceptable o normativamente inválida si concentra beneficios, genera daños distributivos severos o exige capacidades administrativas inexistentes. Por ello, la equidad, la implementabilidad y la estabilidad institucional deben ser modeladas explícitamente.

Sea $\mathcal{G}$ un conjunto finito de grupos relevantes para evaluación distributiva. Para cada $g \in \mathcal{G}$, considérese un funcional lineal del tipo

$$\mathbb{E}[Y_g^{\pi}] = \frac{1}{1-\gamma} \sum_{s,a} d_{\pi}(s, a) y_g(s, a),$$

donde $y_g(s, a)$ representa un resultado de interés del grupo g (cobertura efectiva, ingreso equivalente, tasa de acceso, reducción de tiempos, etc.). Una restricción de brecha media entre grupos g, h puede escribirse como

$$\left| \mathbb{E}[Y_g^\pi] - \mathbb{E}[Y_h^\pi] \right| \leq \delta_{g,h},$$

que equivale a dos desigualdades lineales.

Proposición 5. Si $d \mapsto \mathbb{E}[Y_g^\pi]$ es lineal para todo $g \in \mathcal{G}$, entonces el conjunto de medidas ocupacionales que satisface una familia finita de restricciones de brecha media es convexo.

Demostración. Cada desigualdad linealizada define un semiespacio $a^\top d \leq c$. La intersección finita de semiespacios es convexa. Intersectando con el conjunto de medidas ocupacionales factibles, que es convexo, se preserva convexidad. $\square$

La equidad en medias puede ser insuficiente cuando se requiere protección frente a daños de cola. En ese caso, se introducen restricciones de riesgo sobre pérdidas grupales. Sea L_g^π una pérdida aleatoria asociada al grupo g . La restricción

$$\text{CVaR}_\alpha(L_g^\pi) \leq \bar{\ell}_g$$

controla el promedio de las peores realizaciones por encima del nivel α . La representación convexa estándar por escenarios es

$$\text{CVaR}_\alpha(L) = \min_{\eta \in \mathbb{R}} \left\{ \eta + \frac{1}{1-\alpha} \mathbb{E}[(L - \eta)_+] \right\},$$

lo que permite compatibilidad con formulaciones convexas si la pérdida se expresa en una discretización de escenarios o muestras.

Un punto estructural es la posible incompatibilidad entre restricciones de equidad, presupuesto y capacidad. Para tratarla formalmente, se define

$$\mathcal{D}_{\text{fair}} = \{d \in \mathcal{D}_{\text{base}} : \text{restricciones de equidad}\},$$

donde $\mathcal{D}_{\text{base}}$ denota el conjunto de flujos factibles con restricciones no distributivas. El caso $\mathcal{D}_{\text{fair}} = \emptyset$ no es patológico desde el punto de vista matemático, sino una señal de incompatibilidad normativa-operativa que requiere relajar o penalizar restricciones.

Proposición 6. Si $\mathcal{D}_{\text{fair}} = \emptyset$, una alternativa convexa es la formulación penalizada

$$\max_{\pi \in \Pi} J(\pi) - \lambda_{\text{fair}} \Phi(\pi) \quad \text{sujeta a restricciones no distributivas,}$$

donde $\Phi(\pi) \geq 0$ mide violación agregada de equidad y $\lambda_{\text{fair}} > 0$ controla la severidad de la penalización.

Demostración. La afirmación es constructiva. Si Φ es convexa en d_π y el resto de restricciones define un conjunto convexo no vacío, el problema penalizado es una optimización convexa (o lineal-convexa) bien definida. Además, $\Phi = 0$ en puntos equitativos cuando existen, y $\Phi > 0$ cuantifica violación cuando no existe factibilidad estricta. $\square$

La implementabilidad se modela con un funcional de carga administrativa $h(s, a)$ y un presupuesto de capacidad $\bar{H}$:

$$G_h(\pi) = \frac{1}{1-\gamma} \sum_{s,a} d_\pi(s, a) h(s, a) \leq \bar{H}.$$

En aplicaciones, h puede estimarse por regresión de carga operativa, tiempos de atención, exigencia de sistemas, necesidad de personal o complejidad de coordinación interinstitucional. El punto matemático relevante es que h es un objeto empírico susceptible de error, lo que permite formular sensibilidad.

Proposición 7 (Sensibilidad de implementabilidad). Sean h y $\hat{h}$ dos funciones de carga. Para toda política π ,

$$|G_h(\pi) - G_{\hat{h}}(\pi)| \leq \frac{1}{1-\gamma} \sup_{s,a} |h(s, a) - \hat{h}(s, a)|.$$

Demostración. Por definición y usando que $\sum_{s,a} d_\pi(s, a) = 1$,

$$\begin{aligned} |G_h(\pi) - G_{\hat{h}}(\pi)| &= \frac{1}{1-\gamma} \left| \sum_{s,a} d_\pi(s, a) (h(s, a) - \hat{h}(s, a)) \right| \\ &\leq \frac{1}{1-\gamma} \sum_{s,a} d_\pi(s, a) \sup_{u,v} |h(u, v) - \hat{h}(u, v)|. \end{aligned}$$

Como $\sum_{s,a} d_\pi(s, a) = 1$, se obtiene la cota. $\square$

La estabilidad institucional se puede representar mediante restricciones de proximidad a una política base o a su ocupación d_{π_b} . Una forma linealizable es

$$\sum_{s,a} w_{s,a} |d_{\pi}(s,a) - d_{\pi_b}(s,a)| \leq \Delta,$$

con pesos $w_{s,a} \geq 0$ que capturan sensibilidad operativa o costo político de variar determinadas decisiones. Si los pesos se estiman o fijan normativamente, el modelo conserva auditabilidad al mantenerlos explícitos.

La formulación multiobjetivo se expresa mediante un vector de objetivos

$$\mathbf{F}(\pi) = (F_1(\pi), \dots, F_q(\pi)) \in \mathbb{R}^q.$$

La escalarización lineal

$$J_w(\pi) = \sum_{j=1}^q w_j F_j(\pi), \quad w_j \geq 0, \quad \sum_j w_j = 1,$$

es tractable y transparente, pero sólo recupera puntos soportados de la frontera de Pareto cuando el conjunto de valores alcanzables no es convexo.

Proposición 8. Si el conjunto $\{\mathbf{F}(\pi) : \pi \in \Pi\} \subset \mathbb{R}^q$ es convexo y $w_j > 0$ para todo j , toda solución óptima de $\max_{\pi} J_w(\pi)$ es Pareto-óptima.

Demostración. Supóngase que π^* maximiza J_w pero no es Pareto-óptima. Entonces existe $\tilde{\pi}$ con $F_j(\tilde{\pi}) \geq F_j(\pi^*)$ para todo j y estricta desigualdad para algún j_0 . Multiplicando por $w_j > 0$ y sumando,

$$J_w(\tilde{\pi}) = \sum_j w_j F_j(\tilde{\pi}) > \sum_j w_j F_j(\pi^*) = J_w(\pi^*),$$

contradicción. □

Una alternativa con mayor control normativo es la formulación ϵ -constraint:

$$\max_{\pi} F_1(\pi) \quad \text{s.a.} \quad F_j(\pi) \geq \underline{f}_j, \quad j = 2, \dots, q.$$

Esta forma hace explícitos los mínimos aceptables y evita que un vector de pesos oculte prioridades sustantivas.

La sensibilidad normativa no debe limitarse a la variación del valor de una política fija; también interesa la estabilidad del $\arg \max$. Para una política fija π , la desigualdad

$$|J_w(\pi) - J_{\tilde{w}}(\pi)| = \left| \sum_j (w_j - \tilde{w}_j) F_j(\pi) \right| \leq \|w - \tilde{w}\|_1 \max_j |F_j(\pi)|$$

es inmediata. Sin embargo, el cambio de pesos puede alterar discontinuamente la política óptima si existe empate o margen pequeño.

Teorema 5 (Estabilidad local del selector con margen). *Sea w un vector de pesos y supóngase que existe una política óptima única $\pi^*(w)$ para J_w . Supóngase además que*

$$\Delta(w) := J_w(\pi^*(w)) - \sup_{\pi \neq \pi^*(w)} J_w(\pi) > 0$$

y que $\sup_{\pi \in \Pi} \max_j |F_j(\pi)| \leq M < \infty$. Entonces, para todo $\tilde{w}$ con

$$\|\tilde{w} - w\|_1 < \frac{\Delta(w)}{2M},$$

se mantiene $\pi^*(\tilde{w}) = \pi^*(w)$.

Demostración. Sea $\pi^* = \pi^*(w)$. Para cualquier $\pi \neq \pi^*$,

$$J_w(\pi^*) - J_w(\pi) \geq \Delta(w).$$

Además,

$$|J_{\tilde{w}}(\pi) - J_w(\pi)| \leq M \|\tilde{w} - w\|_1 \quad \forall \pi.$$

Por tanto,

$$J_{\tilde{w}}(\pi^*) - J_{\tilde{w}}(\pi) \geq J_w(\pi^*) - J_w(\pi) - 2M \|\tilde{w} - w\|_1 > \Delta(w) - 2M \cdot \frac{\Delta(w)}{2M} = 0.$$

Se concluye que π^* sigue dominando estrictamente a toda $\pi \neq \pi^*$ bajo $\tilde{w}$, luego $\pi^*(\tilde{w}) = \pi^*(w)$. $\square$

Este resultado formaliza una intuición importante en deliberación pública: si el óptimo tiene margen suficientemente grande, pequeñas variaciones en prioridades normativas no alteran la recomendación. Si el margen es pequeño, la sensibilidad del selector puede ser alta y la IA debe reportar esa fragilidad.

Dimensión	Representación matemática	Propiedad estructural
Equidad de medias	$ \mathbb{E}[Y_g^\pi] - \mathbb{E}[Y_h^\pi] \leq \delta_{g,h}$	Linealizable; convexa en medidas ocupacionales
Riesgo distributivo	$\text{CVaR}_\alpha(L_g^\pi) \leq \bar{\ell}_g$	Convexo por escenarios
Implementabilidad	$G_h(\pi) \leq \bar{H}$	Lineal si h es lineal en d
Estabilidad institucional	$\sum w_{s,a} d - d_b \leq \Delta$	Linealizable con variables auxiliares
Multiobjetivo ponderado	$J_w(\pi) = \sum_j w_j F_j(\pi)$	Tractable; sólo puntos soportados en general
ϵ -constraint	$F_j(\pi) \geq \underline{f}_j$	Umbrales explícitos y auditables
Sensibilidad del sector	$\Delta(w) > 0$ y perturbaciones en w	Estabilidad local con margen

Fuente: Elaboración propia.

6. Causalidad secuencial, identificación y puente formal con el MDP agregado

La optimización descrita en las secciones anteriores requiere como insumos una dinámica causal P , una recompensa causal r y funcionales de restricción g_k interpretables bajo intervención. En entornos públicos, estos objetos deben inferirse a partir de datos y rara vez son observables de manera directa. Sin una capa causal, la IA puede optimizar correlaciones espurias y producir políticas que mejoran métricas históricas sin causar mejoras reales.

Se considera una historia observada hasta tiempo t :

$$H_t = (X_0, A_0, X_1, A_1, \dots, X_t),$$

donde X_t es un vector de covariables observadas y A_t la acción implementada. Para una secuencia $\bar{a}_t = (a_0, \dots, a_t)$, se definen resultados potenciales $X_{t+1}(\bar{a}_t)$ y, en horizonte T , un resultado o recompensa acumulable $Y(\bar{a}_T)$. Se adopta el marco potencial

con supuestos estándar de consistencia, ignorabilidad secuencial y positividad.

Supuesto 1 (Consistencia). Si $\bar{A}_t = \bar{a}_t$, entonces

$$X_{t+1} = X_{t+1}(\bar{a}_t), \quad Y = Y(\bar{a}_T).$$

Supuesto 2 (Ignorabilidad secuencial). Para cada t ,

$$A_t \perp\!\!\!\perp \{X_{t+1}(\bar{a}_t), X_{t+2}(\bar{a}_{t+1}), \dots, Y(\bar{a}_T)\} \mid H_t.$$

Supuesto 3 (Positividad). Si $\mathbb{P}(H_t = h_t) > 0$ y $\pi(a|h_t) > 0$, entonces

$$\mathbb{P}(A_t = a \mid H_t = h_t) > 0.$$

Estos supuestos permiten identificar el valor causal de una política dinámica π , potencialmente historia-dependiente. Para claridad, se trabaja primero con horizonte finito T , con recompensa por período R_t . El valor causal es

$$V_T(\pi) = \mathbb{E} \left[\sum_{t=0}^T \gamma^t R_t^\pi \right].$$

Teorema 6 (Identificación secuencial por g-recursión). *Bajo consistencia, ignorabilidad secuencial y positividad, el valor causal $V_T(\pi)$ es identificable mediante la recursión*

$$V_T(\pi) = \mathbb{E}[m_0^\pi(H_0)],$$

donde

$$m_T^\pi(h_T) = \mathbb{E}[R_T \mid H_T = h_T, A_T \sim \pi(\cdot|h_T)]$$

y, para $t = T - 1, \dots, 0$,

$$m_t^\pi(h_t) = \mathbb{E}[R_t + \gamma m_{t+1}^\pi(H_{t+1}) \mid H_t = h_t, A_t \sim \pi(\cdot|h_t)].$$

Demostración. La prueba se basa en iteración de expectativas. Por definición,

$$V_T(\pi) = \mathbb{E} \left[R_0^\pi + \gamma \sum_{t=1}^T \gamma^{t-1} R_t^\pi \right].$$

Condicionando en H_0 y usando consistencia,

$$V_T(\pi) = \mathbb{E} \left[\mathbb{E} \left[R_0 + \gamma \sum_{t=1}^T \gamma^{t-1} R_t^\pi \mid H_0, A_0 \sim \pi(\cdot | H_0) \right] \right].$$

La ignorabilidad secuencial permite reemplazar distribuciones de potenciales por distribuciones observables condicionales en H_0 . Definiendo

$$m_0^\pi(H_0) = \mathbb{E} [R_0 + \gamma m_1^\pi(H_1) \mid H_0, A_0 \sim \pi(\cdot | H_0)],$$

se obtiene $V_T(\pi) = \mathbb{E}[m_0^\pi(H_0)]$. Iterando el argumento para $t = 1, \dots, T$ se obtiene la recursión. La positividad garantiza que las esperanzas condicionales bajo π están definidas en el soporte de los datos. $\square$

Para conectar esta capa causal con el MDP agregado de optimización, se requiere una aplicación de agregación $S_t = \phi(H_t)$. El punto crítico no es sólo la reducción dimensional, sino la preservación de información relevante para efectos de intervención y restricciones.

Supuesto 4 (Suficiencia causal-Markov de la agregación). Existe una aplicación ϕ tal que $S_t = \phi(H_t)$ satisface, para toda acción a ,

$$\mathbb{P}(S_{t+1} \in B \mid H_t, A_t = a) = \mathbb{P}(S_{t+1} \in B \mid S_t, A_t = a)$$

para todo conjunto $B \subseteq \mathcal{S}$, y

$$\mathbb{E}[R_t \mid H_t, A_t = a] = \mathbb{E}[R_t \mid S_t, A_t = a].$$

Además, las restricciones g_k admiten representación agregada $g_k(S_t, A_t)$ con la misma propiedad de suficiencia para sus esperanzas condicionales.

Proposición 9 (Puente causal-MDP agregado). *Bajo suficiencia causal-Markov de la agregación, las cantidades $P(s' | s, a)$, $r(s, a)$ y $g_k(s, a)$ del MDP agregado representan objetos causales identificables a partir de la historia observada, sujetos a los supuestos de consistencia, ignorabilidad secuencial y positividad.*

Demostración. La suficiencia causal-Markov implica que las distribuciones condicionales y esperanzas relevantes dependen de H_t sólo a través de S_t . Aplicando el

teorema de g-recursión y sustituyendo condicionales en H_t por condicionales en S_t , se obtienen expresiones identificables para transiciones y recompensas agregadas bajo intervención. La positividad asegura soporte para las acciones con probabilidad positiva bajo las políticas a evaluar. $\square$

Esta proposición no elimina la dificultad de validar empíricamente la suficiencia de ϕ , pero ubica con precisión la exigencia matemática que debe satisfacer una agregación si se quiere interpretar la optimización MDP como optimización causal. Si ϕ falla, el optimizador puede seguir siendo internamente coherente respecto al modelo agregado, pero su interpretación causal pierde fundamento.

La formulación causal estándar también enfrenta límites estructurales que no son detalles técnicos menores. La interferencia entre unidades puede violar la descomposición usual de resultados potenciales, las respuestas estratégicas pueden alterar la ley de transición al cambiar la política, y la no estacionariedad puede invalidar transportabilidad de condicionales estimadas. Estos límites se reconocen formalmente porque determinan si el problema admite una extensión del mismo marco o requiere cambio de paradigma (por ejemplo juegos dinámicos, modelos de red o equilibrio general).

A nivel de una etapa, la g-fórmula estática se recupera como caso particular. Bajo consistencia, ignorabilidad condicional y positividad,

$$\mathbb{E}[Y(a)] = \mathbb{E}[\mathbb{E}[Y \mid A = a, X]].$$

El caso dinámico es estrictamente más exigente porque la acción altera covariables futuras y éstas condicionan decisiones posteriores. Por esa razón, la identificación secuencial no es un lujo metodológico, sino una condición mínima cuando se optimizan políticas persistentes o adaptativas.

Componente causal	Nivel	Consecuencia si falla
Consistencia	Definición de intervención	Ambigüedad entre observado y potencial
Ignorabilidad secuencial	Identificación dinámica	Sesgo por confusión temporal no controlada
Positividad	Soporte	Imposibilidad de evaluar ciertas políticas con datos disponibles

Componente causal	Nivel	Consecuencia si falla
Suficiencia de agregación ϕ	Puente causal-MDP	Optimización sobre un estado no causalmente suficiente
Ausencia de interferencia (o controlada)	Estructura de unidades	g-recursión estándar puede dejar de ser válida
Estabilidad estructural / transportabilidad	Temporal o contextual	Fallo de extrapolación de efectos estimados

Fuente: Elaboración propia.

La sección siguiente incorpora evaluación fuera de política y cotas de error de modelo, que enlazan la capa causal-estadística con la capa de optimización y permiten cuantificar el riesgo de optimizar sobre modelos estimados en lugar del sistema real.

7. Evaluación fuera de política, doble robustez, clipping, cotas de error y suboptimalidad bajo modelo estimado

En administración pública, el despliegue completo de una nueva política suele ser costoso, lento o políticamente sensible. Por ello, una IA pública debe poder evaluar políticas candidatas usando datos recolectados bajo una política de comportamiento histórica β . Este problema es la evaluación fuera de política (*off-policy evaluation*, OPE). Las razones de importancia ofrecen una identidad exacta bajo soporte común, pero su varianza puede ser alta; los estimadores doblemente robustos reducen fragilidad al combinar reponderación con modelos de resultados.

Considérese un horizonte finito T , una trayectoria

$$\tau = (s_0, a_0, r_0, \dots, s_T, a_T, r_T, s_{T+1}),$$

y retorno descontado

$$G_T(\tau) = \sum_{t=0}^T \gamma^t r_t.$$

La razón de importancia de trayectoria es

$$W_T(\tau) = \prod_{t=0}^T \frac{\pi(a_t|s_t)}{\beta(a_t|s_t)}.$$

Se asume soporte común:

$$\beta(a|s) > 0 \quad \text{si} \quad \pi(a|s) > 0.$$

Teorema 7 (Insesgo del estimador de importancia). Si $\tau_1, \dots, \tau_n$ son trayectorias i.i.d. generadas bajo β , entonces

$$\hat{J}_{\text{IS}} = \frac{1}{n} \sum_{i=1}^n W_T(\tau_i) G_T(\tau_i)$$

es insesgado para $J_T(\pi) = \mathbb{E}_\pi[G_T]$.

Demostración. La probabilidad de una trayectoria bajo β es

$$p_\beta(\tau) = \mu(s_0) \prod_{t=0}^T \beta(a_t|s_t) P(s_{t+1}|s_t, a_t).$$

Multiplicando por $W_T(\tau)$,

$$W_T(\tau) p_\beta(\tau) = \mu(s_0) \prod_{t=0}^T \pi(a_t|s_t) P(s_{t+1}|s_t, a_t) = p_\pi(\tau).$$

Por tanto,

$$\mathbb{E}_\beta[W_T G_T] = \sum_{\tau} W_T(\tau) G_T(\tau) p_\beta(\tau) = \sum_{\tau} G_T(\tau) p_\pi(\tau) = \mathbb{E}_\pi[G_T].$$

Promediar n copias preserva el insesgo. □

El insesgo de IS no controla la varianza. Cuando π y β difieren sustancialmente, W_T puede tener colas pesadas. Una práctica frecuente es usar clipping de razones de importancia para reducir varianza a costa de introducir sesgo.

Proposición 10 (Sesgo por clipping: cota simple). Sea $\tilde{W}_T = \min\{W_T, c\}$ con $c > 0$ y

$$\hat{J}_{\text{CIS}} = \frac{1}{n} \sum_{i=1}^n \tilde{W}_T(\tau_i) G_T(\tau_i).$$

Si $|G_T| \leq G_{\text{máx}}$ casi seguramente, entonces

$$|\mathbb{E}_\beta[\hat{J}_{\text{CIS}}] - J_T(\pi)| \leq G_{\text{máx}} \mathbb{E}_\beta[(W_T - c)_+].$$

Demostración. Por linealidad de la esperanza,

$$\mathbb{E}_\beta[\tilde{W}_T G_T] - \mathbb{E}_\beta[W_T G_T] = \mathbb{E}_\beta[(\tilde{W}_T - W_T) G_T].$$

Como $\tilde{W}_T - W_T = -(W_T - c)_+$,

$$|\mathbb{E}_\beta[\tilde{W}_T G_T] - J_T(\pi)| \leq \mathbb{E}_\beta[(W_T - c)_+ |G_T|] \leq G_{\text{máx}} \mathbb{E}_\beta[(W_T - c)_+].$$

□

Esta cota no pretende ser ajustada, pero formaliza el intercambio varianza-sesgo y justifica reportar explícitamente el nivel de clipping y el sesgo potencial inducido.

Se introduce ahora un estimador doblemente robusto en forma estándar para horizonte finito. Sea H_t la historia hasta t , $\rho_{0:t} = \prod_{u=0}^t \pi(A_u | H_u) / \beta(A_u | H_u)$, sea $\hat{Q}_t(h_t, a)$ una aproximación del valor-acción y

$$\hat{V}_t(h_t) = \sum_a \pi(a | h_t) \hat{Q}_t(h_t, a).$$

Considérese

$$\hat{J}_{\text{DR}} = \frac{1}{n} \sum_{i=1}^n \left[\hat{V}_0(H_0^i) + \sum_{t=0}^T \gamma^t \rho_{0:t}^i \left(R_t^i + \gamma \hat{V}_{t+1}(H_{t+1}^i) - \hat{Q}_t(H_t^i, A_t^i) \right) \right].$$

Teorema 8 (Doble robustez, resultado estándar con prueba esquemática). *Bajo positividad, integrabilidad y regularidad suficiente para ley de grandes números, $\hat{J}_{\text{DR}}$ es consistente para $J_T(\pi)$ si se cumple al menos una de las siguientes condiciones:*

- 1) las razones de importancia usan la política de comportamiento correcta β ;

2) los modelos de resultado convergen a los verdaderos Q_t^π y V_t^π .

Demostración. La demostración estándar se basa en una descomposición telescópica de los residuos

$$R_t + \gamma \hat{V}_{t+1}(H_{t+1}) - \hat{Q}_t(H_t, A_t).$$

Si β es correcta, la reponderación transporta expectativas desde la distribución de comportamiento a la distribución de evaluación; la suma telescópica cancela los errores de aproximación en $\hat{Q}_t, \hat{V}_t$ en el límite de medias. Si $\hat{Q}_t, \hat{V}_t$ son correctos, cada residual tiene esperanza condicional nula y la suma colapsa al valor verdadero aun si la parte de reponderación es imperfecta en la parametrización intermedia, siempre que el término implementado converja a una forma equivalente. Bajo condiciones de integrabilidad y convergencia de medias, se obtiene consistencia. $\square$

El resultado anterior se incluye como teorema estándar porque una demostración completa con todas las hipótesis de regularidad y detalles de recursión excede el alcance de este texto. La utilidad práctica reside en la estructura de tolerancia a error parcial del módulo causal-estadístico, especialmente en entornos con datos administrativos heterogéneos.

Más allá de OPE, la IA pública necesita cuantificar el error de optimizar sobre un modelo estimado $\hat{M} = (\hat{P}, \hat{r})$ en lugar del modelo verdadero $M = (P, r)$. Para una política fija, una cota tipo *simulation lemma* controla la diferencia de valores. Se define

$$\begin{aligned} R_{\max} &= \sup_{s,a} |r(s,a)|, \\ \varepsilon_r &= \sup_{s,a} |\hat{r}(s,a) - r(s,a)|, \\ \varepsilon_P &= \sup_{s,a} \text{TV}(\hat{P}(\cdot|s,a), P(\cdot|s,a)). \end{aligned}$$

Teorema 9 (Cota de error de valor por mala especificación del modelo). *Para cualquier política π ,*

$$\|V_{\hat{M}}^\pi - V_M^\pi\|_\infty \leq \frac{\varepsilon_r}{1-\gamma} + \frac{2\gamma R_{\max}}{(1-\gamma)^2} \varepsilon_P.$$

Demostración. Usando las ecuaciones de Bellman bajo $\hat{M}$ y M ,

$$V_{\hat{M}}^\pi - V_M^\pi = T_{\hat{M}}^\pi V_{\hat{M}}^\pi - T_M^\pi V_M^\pi = (T_{\hat{M}}^\pi V_{\hat{M}}^\pi - T_{\hat{M}}^\pi V_M^\pi) + (T_{\hat{M}}^\pi V_M^\pi - T_M^\pi V_M^\pi).$$

Tomando norma supremo y aplicando contracción de $T_{\hat{M}}^\pi$,

$$\|V_{\hat{M}}^\pi - V_M^\pi\|_\infty \leq \gamma \|V_{\hat{M}}^\pi - V_M^\pi\|_\infty + \|(T_{\hat{M}}^\pi - T_M^\pi)V_M^\pi\|_\infty.$$

Luego,

$$(1 - \gamma) \|V_{\hat{M}}^\pi - V_M^\pi\|_\infty \leq \|(T_{\hat{M}}^\pi - T_M^\pi)V_M^\pi\|_\infty.$$

Para cada s ,

$$\begin{aligned} \left| ((T_{\hat{M}}^\pi - T_M^\pi)V_M^\pi)(s) \right| &\leq \sum_a \pi(a|s) \left(|\hat{r}(s, a) - r(s, a)| \right. \\ &\quad \left. + \gamma \left| \sum_{s'} (\hat{P} - P)(s'|s, a) V_M^\pi(s') \right| \right). \end{aligned}$$

Usando $|\hat{r} - r| \leq \varepsilon_r$ y la desigualdad

$$\left| \sum_x (p - q)(x) f(x) \right| \leq 2 \text{TV}(p, q) \|f\|_\infty,$$

se obtiene

$$\|(T_{\hat{M}}^\pi - T_M^\pi)V_M^\pi\|_\infty \leq \varepsilon_r + 2\gamma\varepsilon_P \|V_M^\pi\|_\infty.$$

Como $\|V_M^\pi\|_\infty \leq R_{\text{máx}}/(1 - \gamma)$, resulta

$$\|(T_{\hat{M}}^\pi - T_M^\pi)V_M^\pi\|_\infty \leq \varepsilon_r + \frac{2\gamma R_{\text{máx}}}{1 - \gamma} \varepsilon_P.$$

Dividiendo por $1 - \gamma$, se obtiene la cota. □

La cota para una política fija es útil, pero la IA selecciona una política $\hat{\pi}$ optimizando sobre $\hat{M}$. Interesa entonces una cota de suboptimalidad en el modelo verdadero.

Teorema 10 (Cota de suboptimalidad de la política optimizada en $\hat{M}$). *Sea $\hat{\pi} \in \arg \max_{\pi} J_{\hat{M}}(\pi)$ y $\pi^* \in \arg \max_{\pi} J_M(\pi)$. Supóngase que*

$$\sup_{\pi \in \Pi} |J_{\hat{M}}(\pi) - J_M(\pi)| \leq \Delta_M.$$

Entonces

$$J_M(\pi^*) - J_M(\hat{\pi}) \leq 2\Delta_M.$$

Demostración. Por optimalidad de $\hat{\pi}$ en $\hat{M}$,

$$J_{\hat{M}}(\hat{\pi}) \geq J_{\hat{M}}(\pi^*).$$

Por la cota uniforme,

$$J_M(\hat{\pi}) \geq J_{\hat{M}}(\hat{\pi}) - \Delta_M \geq J_{\hat{M}}(\pi^*) - \Delta_M \geq J_M(\pi^*) - 2\Delta_M.$$

Reordenando,

$$J_M(\pi^*) - J_M(\hat{\pi}) \leq 2\Delta_M.$$

□

Combinando este teorema con cotas uniformes derivadas de errores en P y r se obtiene una garantía explícita de degradación máxima por optimizar sobre un modelo estimado. En contextos públicos, esta degradación es relevante porque permite convertir incertidumbre estadística en márgenes de seguridad o en penalizaciones conservadoras durante el despliegue.

Herramienta	Garantía principal	Riesgo / costo
IS	Insesgo exacto bajo soporte común	Varianza alta cuando π y β divergen
Clipping de IS	Reduce varianza práctica	Introduce sesgo (cuantificable por colas de W_T)
DR	Consistencia con corrección parcial	Requiere modelos auxiliares e hipótesis adicionales
Cota de valor por modelo	Control de sensibilidad para política fija	Conservadora en horizontes largos $((1 - \gamma)^{-2})$
Cota de suboptimalidad	Control de pérdida por optimizar en $\hat{M}$	Requiere cota uniforme sobre $J_{\hat{M}} - J_M$

Fuente: Elaboración propia.

La siguiente sección incorpora robustez frente a incertidumbre de transición, no estacionariedad y seguridad probabilística respecto a una política base, integrando estos elementos con el lema de diferencia de desempeño y con regularización por divergencia.

8. Robustez, no estacionariedad, seguridad probabilística, estabilidad por KL y aprendizaje primal-dual

La estimación causal de P y r no elimina incertidumbre. En políticas públicas, shocks macroeconómicos, cambios regulatorios, aprendizaje de agentes, crisis y conflictos pueden generar desviaciones sistemáticas entre la dinámica estimada y la realizada. Para tratar esta fragilidad se introduce un MDP robusto con incertidumbre rectangular, se formaliza seguridad respecto a una política base mediante restricciones probabilísticas y se discute convergencia primal-dual en el régimen estándar de aproximación estocástica.

Para cada (s, a) , se define un conjunto de incertidumbre de transiciones $\mathcal{P}_{s,a}$, no vacío, compacto y convexo. Bajo rectangularidad por pares (s, a) , el operador robusto de Bellman para una política π es

$$(T_R^\pi V)(s) = \sum_a \pi(a|s) \left[r(s, a) + \gamma \inf_{p \in \mathcal{P}_{s,a}} \sum_{s'} p(s') V(s') \right].$$

Supuesto 5 (Rectangularidad y compacidad). Para todo (s, a) , $\mathcal{P}_{s,a} \subset \Delta(\mathcal{S})$ es no vacío, compacto y convexo, y la incertidumbre es (s, a) -rectangular.

Lema 2. Para toda política π , el operador T_R^π es una contracción en norma supremo con módulo γ :

$$\|T_R^\pi V - T_R^\pi W\|_\infty \leq \gamma \|V - W\|_\infty.$$

Demostración. Defínase $\phi_{s,a}(V) = \inf_{p \in \mathcal{P}_{s,a}} \sum_{s'} p(s') V(s')$. Para cualquier $p \in \mathcal{P}_{s,a}$,

$$\sum_{s'} p(s') V(s') \leq \sum_{s'} p(s') W(s') + \|V - W\|_\infty.$$

Tomando ínfimo en p ,

$$\phi_{s,a}(V) \leq \phi_{s,a}(W) + \|V - W\|_\infty.$$

Intercambiando V, W , se obtiene

$$|\phi_{s,a}(V) - \phi_{s,a}(W)| \leq \|V - W\|_\infty.$$

Luego, para cada s ,

$$|(T_R^\pi V)(s) - (T_R^\pi W)(s)| \leq \sum_a \pi(a|s) \gamma |\phi_{s,a}(V) - \phi_{s,a}(W)| \leq \gamma \|V - W\|_\infty.$$

Tomando supremo, queda probado. $\square$

Corolario 3. *Existe un único V_R^π tal que $V_R^\pi = T_R^\pi V_R^\pi$.*

El operador robusto óptimo se define por

$$(T_R^* V)(s) = \max_{a \in \mathcal{A}} \left[r(s, a) + \gamma \inf_{p \in \mathcal{P}_{s,a}} \sum_{s'} p(s') V(s') \right].$$

Teorema 11 (Optimalidad robusta bajo rectangularidad). *Bajo el supuesto de rectangularidad y compacidad, T_R^* es una contracción con módulo γ , posee un único punto fijo V_R^* , y toda política codiciosa con respecto a V_R^* es robustamente óptima.*

Demostración. La contracción se demuestra igual que en el lema anterior, sustituyendo promedio por máximo. La existencia y unicidad del punto fijo se obtiene por el teorema de Banach. La optimalidad de una política codiciosa se sigue del argumento estándar de programación dinámica aplicado al operador robusto, cuya consistencia dinámica está garantizada por la rectangularidad. $\square$

La calibración de $\mathcal{P}_{s,a}$ es un problema sustantivo. El marco admite conjuntos basados en distancia TV, KL, χ^2 , Wasserstein, o intervalos de confianza multinomiales, siempre que respeten la estructura rectangular si se desea preservar el principio de optimalidad. La matemática separa aquí dos cuestiones: la validez de la programación robusta dada una familia $\mathcal{P}_{s,a}$, y la adecuación empírica de esa familia a la incertidumbre real.

Proposición 11 (Monotonicidad respecto al tamaño del conjunto robusto). *Supóngase que para todo (s, a) , $\mathcal{P}_{s,a}^{(1)} \subseteq \mathcal{P}_{s,a}^{(2)}$. Denote por $V_{R,1}^*$ y $V_{R,2}^*$ los valores robustamente óptimos correspondientes. Entonces*

$$V_{R,1}^*(s) \geq V_{R,2}^*(s) \quad \forall s \in \mathcal{S}.$$

Demostración. Para cualquier V y (s, a) ,

$$\inf_{p \in \mathcal{P}_{s,a}^{(1)}} \sum_{s'} p(s') V(s') \geq \inf_{p \in \mathcal{P}_{s,a}^{(2)}} \sum_{s'} p(s') V(s').$$

Por tanto, $T_{R,1}^* V \geq T_{R,2}^* V$ coordenada a coordenada para todo V . Iterando desde un mismo vector inicial y usando convergencia por contracción de ambos operadores a sus puntos fijos, se concluye $V_{R,1}^* \geq V_{R,2}^*$. $\square$

Esta proposición formaliza que robustecer con conjuntos más grandes reduce (o no aumenta) el valor robusto, lo cual es coherente con la interpretación conservadora de incertidumbre adversarial.

La no estacionariedad se incorpora en forma explícita con horizonte finito y dinámica dependiente del tiempo:

$$P_t(s'|s, a), \quad r_t(s, a), \quad \pi_t(a|s), \quad t = 0, \dots, T.$$

La recursión de Bellman se vuelve

$$V_t^\pi(s) = \sum_a \pi_t(a|s) \left[r_t(s, a) + \gamma \sum_{s'} P_t(s'|s, a) V_{t+1}^\pi(s') \right], \quad V_{T+1}^\pi \equiv 0.$$

Una forma simple de cuantificar drift temporal en transiciones es

$$\delta_t = \sup_{s,a} \text{TV}(P_t(\cdot|s, a), P_{t+1}(\cdot|s, a)).$$

Si δ_t es pequeña y acumulada de manera controlada, las políticas recalibradas periódicamente pueden mantener garantías aproximadas; si el drift es grande o endógeno a la propia política, pueden requerirse marcos adaptativos más complejos.

La seguridad respecto a una política base π_b es una condición normativa y operativa central. Una versión determinista en términos de valores verdaderos es

$$J(\pi) \geq J(\pi_b) - \varepsilon.$$

Sin embargo, en despliegue real J es desconocido y debe ser estimado. Por ello, se introduce una formulación probabilística con cotas de confianza.

Definición 5. Se dirá que una política π satisface *seguridad probabilística* frente a π_b al nivel $1 - \alpha$ si

$$\mathbb{P}(J(\pi) \geq J(\pi_b) - \varepsilon) \geq 1 - \alpha.$$

Sea $\hat{J}(\pi)$ un estimador de $J(\pi)$ y sea $\text{LCB}_\alpha(\pi)$ una cota inferior tal que

$$\mathbb{P}(J(\pi) \geq \hat{J}(\pi) - \text{LCB}_\alpha(\pi)) \geq 1 - \alpha/2,$$

y $\text{UCB}_\alpha(\pi_b)$ una cota superior de error para π_b tal que

$$\mathbb{P}(J(\pi_b) \leq \hat{J}(\pi_b) + \text{UCB}_\alpha(\pi_b)) \geq 1 - \alpha/2.$$

Proposición 12 (Regla suficiente de despliegue seguro). *Si una política π satisface la condición observable*

$$\hat{J}(\pi) - \text{LCB}_\alpha(\pi) \geq \hat{J}(\pi_b) + \text{UCB}_\alpha(\pi_b) - \varepsilon,$$

entonces

$$\mathbb{P}(J(\pi) \geq J(\pi_b) - \varepsilon) \geq 1 - \alpha,$$

suponiendo validez simultánea de las cotas con probabilidad al menos $1 - \alpha$ por unión.

Demostración. Por la validez de las cotas,

$$J(\pi) \geq \hat{J}(\pi) - \text{LCB}_\alpha(\pi), \quad J(\pi_b) \leq \hat{J}(\pi_b) + \text{UCB}_\alpha(\pi_b)$$

con probabilidad al menos $1 - \alpha$ (por unión de eventos con nivel $\alpha/2$ cada uno). En ese evento,

$$J(\pi) \geq \hat{J}(\pi) - \text{LCB}_\alpha(\pi) \geq \hat{J}(\pi_b) + \text{UCB}_\alpha(\pi_b) - \varepsilon \geq J(\pi_b) - \varepsilon.$$

□

El lema de diferencia de desempeño permite formular seguridad también en términos de ventajas respecto a π_b :

$$J(\pi) - J(\pi_b) = \frac{1}{1 - \gamma} \mathbb{E}_{d_\pi}[A^{\pi_b}(s, a)].$$

Así, una condición suficiente determinista es

$$\mathbb{E}_{d_\pi}[A^{\pi_b}(s, a)] \geq -(1 - \gamma)\varepsilon.$$

En práctica, la dificultad radica en estimar A^{π_b} y d_π ; por ello, la versión probabilística con cotas observables es preferible para despliegue.

Para estabilizar cambios respecto a la política base y reducir varianza de OPE, se usa regularización por divergencia KL. Sea ν una distribución de referencia sobre estados y supóngase soporte común o suavización de la política base. La penalización toma la forma

$$\max_{\pi} J(\pi) - \tau \mathbb{E}_{s \sim \nu} [\text{KL}(\pi(\cdot|s) \parallel \pi_b^\varepsilon(\cdot|s))],$$

donde

$$\pi_b^\varepsilon(\cdot|s) = (1 - \varepsilon)\pi_b(\cdot|s) + \varepsilon u(\cdot|s), \quad \varepsilon \in (0, 1),$$

y $u(\cdot|s)$ es una distribución de soporte completo.

Proposición 13. Para ν , π_b^ε y $\tau > 0$ fijos, la aplicación

$$\pi \mapsto -\tau \mathbb{E}_{s \sim \nu} [\text{KL}(\pi(\cdot|s) \parallel \pi_b^\varepsilon(\cdot|s))]$$

es cóncava en π .

Demostración. La divergencia $\text{KL}(\cdot \parallel \pi_b^\varepsilon(\cdot|s))$ es convexa en su primer argumento para cada s . Multiplicar por $-\tau$ preserva cóncavidad, y la esperanza ponderada por ν también. $\square$

Se considera el aprendizaje primal-dual estocástico para parametrizaciones π_θ , con lagrangiano

$$\mathcal{L}(\theta, \lambda) = J(\pi_\theta) - \sum_{k=1}^m \lambda_k (G_k(\pi_\theta) - b_k), \quad \lambda \in \mathbb{R}_+^m.$$

Un esquema básico es

$$\theta_{n+1} = \theta_n + \eta_n \nabla_{\theta} \widehat{\mathcal{L}}(\theta_n, \lambda_n), \quad \lambda_{n+1} = \Pi_{\mathbb{R}_+^m} \left(\lambda_n + \rho_n (\widehat{G}(\theta_n) - b) \right).$$

Supuesto 6 (Pasos de Robbins–Monro en dos escalas). Las sucesiones (η_n) y (ρ_n)

satisfacen

$$\sum_n \eta_n = \infty, \quad \sum_n \eta_n^2 < \infty, \quad \sum_n \rho_n = \infty, \quad \sum_n \rho_n^2 < \infty,$$

y existe separación de escalas, por ejemplo $\rho_n/\eta_n \rightarrow 0$ o $\eta_n/\rho_n \rightarrow 0$.

Supuesto 7 (Ruido de gradiente y regularidad). Los estimadores de gradiente son diferencias de martingala con segundo momento uniformemente acotado, $\mathcal{L}$ es suficientemente suave (Lipschitz de gradientes en regiones relevantes), y los iterados permanecen en una región compacta o efectivamente compacta.

Teorema 12 (Convergencia primal-dual estocástica, resultado estándar). *Bajo los supuestos anteriores, las trayectorias interpoladas del esquema primal-dual estocástico son asintóticamente cercanas a la dinámica ODE proyectada asociada, y los iterados convergen casi seguramente a un conjunto internamente invariante; en el caso convexo-concavo regular, el límite pertenece al conjunto de puntos KKT, y en el caso no convexo el límite pertenece al conjunto de puntos estacionarios de la dinámica proyectada.*

Demostración. La prueba es estándar en aproximación estocástica con dos escalas. La proyección mantiene factibilidad dual, las condiciones de Robbins–Monro y el control de ruido permiten aplicar el método ODE, y la separación de escalas induce una dinámica cuasi-estática para la variable lenta respecto de la rápida. En el caso convexo-concavo, la teoría primal-dual estándar identifica el conjunto de equilibrios con KKT; en no convexidad, la conclusión se formula en términos de conjuntos internamente invariantes/estacionarios de la dinámica proyectada. $\square$

Módulo	Garantía matemática	Supuestos críticos
Robust MDP	Punto fijo y política robusta codiciosa	Rectangularidad, compacidad, $\gamma < 1$
Monotonicidad robusta	Conjuntos más grandes $\Rightarrow$ valor robusto menor	Inclusión $\mathcal{P}^{(1)} \subseteq \mathcal{P}^{(2)}$
No estacionariedad finita	Bellman por inducción hacia atrás	Especificación temporal P_t, r_t
Seguridad probabilística	No deterioro con probabilidad $\geq 1 - \alpha$	Cotas válidas de inferencia para π, π_b
Regularización por KL	Concavidad del término de estabilidad	Soporte común o suavización π_b^ϵ

Módulo	Garantía matemática	Supuestos críticos
Primal-dual estocástico	Convergencia estándar a invariantes/KKT	RM, ruido controlado, suavidad, compacidad efectiva

Fuente: Elaboración propia.

9. Caso estático de asignación presupuestaria, complejidad computacional, límites estructurales y síntesis

Como bloque analítico cerrable, se incluye un problema estático de asignación presupuestaria con rendimientos decrecientes y solución cerrada. Este caso sirve como validación interna de la arquitectura: muestra que el marco no depende de manera esencial de métodos numéricos opacos y que conserva interpretabilidad económica incluso en módulos simplificados.

Sea $x_i \geq 0$ el gasto asignado al programa $i = 1, \dots, n$. Se modela la utilidad social de cada programa como

$$u_i(x_i) = a_i \log(1 + b_i x_i), \quad a_i, b_i > 0.$$

El problema es

$$\max_{x \in \mathbb{R}_+^n} \sum_{i=1}^n a_i \log(1 + b_i x_i) \quad \text{s.a.} \quad \sum_{i=1}^n x_i \leq B.$$

Teorema 13. *El problema de asignación presupuestaria anterior tiene solución óptima única.*

Demostración. Cada u_i es continua y estrictamente cóncava en $\mathbb{R}_+$, porque

$$u_i''(x_i) = -\frac{a_i b_i^2}{(1 + b_i x_i)^2} < 0.$$

La suma $\sum_i u_i(x_i)$ es estrictamente cóncava. El conjunto factible $\{x \in \mathbb{R}_+^n : \sum_i x_i \leq B\}$ es no vacío, cerrado y acotado, luego compacto. Por Weierstrass existe solución; por estricta concavidad en un conjunto convexo, la solución es única. $\square$

Teorema 14 (Solución cerrada tipo *water-filling*). Existe $\lambda^* \geq 0$ tal que la solución óptima única satisface

$$x_i^* = \max \left\{ 0, \frac{a_i}{\lambda^*} - \frac{1}{b_i} \right\}, \quad i = 1, \dots, n,$$

y además $\sum_i x_i^* = B$.

Demostración. El lagrangiano con multiplicadores $\lambda \geq 0$ para presupuesto y $v_i \geq 0$ para no negatividad es

$$\mathcal{L}(x, \lambda, \nu) = \sum_i a_i \log(1 + b_i x_i) - \lambda \left(\sum_i x_i - B \right) + \sum_i v_i x_i.$$

La estacionariedad exige

$$\frac{a_i b_i}{1 + b_i x_i} - \lambda + v_i = 0.$$

Si $x_i^* > 0$, por complementariedad $v_i^* = 0$, luego

$$\frac{a_i b_i}{1 + b_i x_i^*} = \lambda^* \implies x_i^* = \frac{a_i}{\lambda^*} - \frac{1}{b_i}.$$

Si este valor es negativo, la restricción $x_i \geq 0$ se activa y $x_i^* = 0$. En conjunto,

$$x_i^* = \max \left\{ 0, \frac{a_i}{\lambda^*} - \frac{1}{b_i} \right\}.$$

Resta probar que el presupuesto es vinculante. Supóngase por contradicción que $\sum_i x_i^* < B$. Entonces, por complementariedad de la restricción presupuestaria, $\lambda^* = 0$.

La estacionariedad implica

$$\frac{a_i b_i}{1 + b_i x_i^*} + v_i^* = 0.$$

Pero $\frac{a_i b_i}{1 + b_i x_i^*} > 0$ y $v_i^* \geq 0$, lo que es imposible. Equivalente y operativamente, si hubiera holgura presupuestaria, la derivada marginal de algún programa sería positiva y podría aumentarse su gasto mejorando el objetivo. Por tanto, $\sum_i x_i^* = B$. $\square$

Este bloque muestra cómo los precios sombra aparecen en una forma analítica simple. El multiplicador λ^* iguala utilidades marginales ajustadas para programas activos y excluye aquellos cuya productividad marginal inicial $a_i b_i$ no justifica asignación positiva bajo la restricción de presupuesto. En contextos públicos, esta lógica permite

interpretar decisiones de priorización sin recurrir a cajas negras.

La escalabilidad computacional merece tratamiento explícito, porque el título del documento incluye IA y el núcleo exacto del texto usa MDP finitos. En espacios de estado grandes, la representación exacta por medidas ocupacionales puede ser inviable. Si $|S| = S$ y $|A| = A$, el primal exacto involucra SA variables básicas (más auxiliares para linealizaciones) y al menos $S + m$ restricciones principales, lo que crece rápidamente con discretizaciones finas.

Para enfrentar esta limitación, se suele recurrir a aproximación funcional, agregación de estados, muestreo o descomposición. Matemáticamente, ello transforma garantías exactas en garantías aproximadas. Una manera abstracta de capturar esta transición es introducir un operador de proyección $\Pi_{\mathcal{F}}$ sobre una clase de funciones/medidas aproximantes $\mathcal{F}$, de modo que la ecuación de Bellman o el problema primal se resuelvan en una subclase restringida. El error total se descompone entonces en error de aproximación, error estadístico y error de optimización numérica.

Proposición 14 (Descomposición abstracta de error con aproximación funcional). *Sea π^* una política óptima en la clase exacta Π y $\hat{\pi}_{\mathcal{F}}$ una política obtenida al optimizar sobre una clase aproximante $\Pi_{\mathcal{F}} \subseteq \Pi$ con modelo estimado. Entonces, para todo entorno M ,*

$$\begin{aligned} J_M(\pi^*) - J_M(\hat{\pi}_{\mathcal{F}}) &\leq \underbrace{\inf_{\pi \in \Pi_{\mathcal{F}}} (J_M(\pi^*) - J_M(\pi))}_{\text{error de aproximación}} \\ &\quad + \underbrace{\sup_{\pi \in \Pi_{\mathcal{F}}} |J_M(\pi) - J_{\hat{M}}(\pi)|}_{\text{error estadístico/modelo}} \\ &\quad + \underbrace{\left(\sup_{\pi \in \Pi_{\mathcal{F}}} J_{\hat{M}}(\pi) - J_{\hat{M}}(\hat{\pi}_{\mathcal{F}}) \right)}_{\text{error de optimización numérica}}. \end{aligned}$$

Demostración. Para cualquier $\tilde{\pi} \in \Pi_{\mathcal{F}}$,

$$\begin{aligned} J_M(\pi^*) - J_M(\hat{\pi}_{\mathcal{F}}) &= (J_M(\pi^*) - J_M(\tilde{\pi})) + (J_M(\tilde{\pi}) - J_{\hat{M}}(\tilde{\pi})) \\ &\quad + (J_{\hat{M}}(\tilde{\pi}) - J_{\hat{M}}(\hat{\pi}_{\mathcal{F}})) \\ &\quad + (J_{\hat{M}}(\hat{\pi}_{\mathcal{F}}) - J_M(\hat{\pi}_{\mathcal{F}})). \end{aligned}$$

Acotando los dos términos de desajuste M vs $\hat{M}$ por el supremo uniforme en $\Pi_{\mathcal{F}}$, y

luego tomando ínfimo en $\tilde{\pi} \in \Pi_{\mathcal{F}}$, se obtiene la desigualdad. □

Esta proposición reafirma que “IA” no debe leerse como sinónimo de aprendizaje automático no restringido, sino como un esquema donde la complejidad computacional se gestiona mediante aproximación, sin perder trazabilidad de fuentes de error.

El marco desarrollado tiene límites estructurales que conviene clasificar. Algunos son extensiones naturales dentro del mismo paradigma (por ejemplo no estacionariedad finita, restricciones adicionales convexas, robustez local, mejoras de inferencia en OPE). Otros exigen cambiar la estructura del modelo (interferencia fuerte entre unidades, juegos dinámicos estratégicos, equilibrio general con precios endógenos, observabilidad parcial severa con memoria larga no agregable, o instituciones que alteran el espacio de acciones de forma no exógena). Esta clasificación evita sobreextender interpretaciones de los teoremas probados.

Límite o extensión	Tipo	Implicación formal
No estacionariedad de horizonte finito	Extensión natural	Bellman temporal; no cambia la lógica base de DP
Restricciones convexas adicionales	Extensión natural	Se preserva tractabilidad con LP/convexo ampliado
Robustez local de transiciones	Extensión natural	MDP robusto con rectangularidad
Interferencia fuerte entre unidades	Cambio de marco	g-recursión estándar y MDP agregado pueden fallar
Juegos dinámicos estratégicos	Cambio de marco	Requiere equilibrio (Markov perfect, etc.)
Equilibrio general con precios endógenos	Cambio de marco	La transición ya no es exógena al policy maker
POMDP severo no agregable	Cambio de marco	Estado observable no suficiente; requiere creencias/filtrado

Fuente: Elaboración propia.

La síntesis estructural del documento puede expresarse como una cadena de objetos matemáticos:

$$(\mathcal{D}_n, \mathcal{C}) \xrightarrow{\hat{\Psi}_n} (\hat{P}, \hat{r}, \hat{g}, \hat{\mathcal{U}}, \hat{\mathcal{I}}) \xrightarrow{\mathcal{O}} \hat{\pi} \xrightarrow{\text{seguridad/robustez}} \pi^{\text{dep}} \xrightarrow{\text{monitoreo}} \mathcal{D}_{n+1}.$$

La última flecha hace explícita la realimentación de datos y el papel del monitoreo de drift. Matemáticamente, esto cierra el ciclo adaptativo de una IA pública sin suponer que el problema termina en la optimización inicial.

El aporte del marco no reside en afirmar neutralidad normativa automática ni en prometer validez fuera de supuestos. Su valor está en convertir objetivos, restricciones, supuestos causales, incertidumbre y criterios de seguridad en objetos formales trazables. Esa explicitud permite demostrar resultados, identificar puntos de fragilidad, cuantificar sensibilidad y someter decisiones a auditoría técnica y deliberación pública informada.

Componente del sistema	Objeto matemático	Papel funcional en IA pública
Objetivos sociales	$J(\pi), F(\pi)$, pesos w, α	Define qué se optimiza y con qué prioridades explícitas
Restricciones públicas	$G_k(\pi) \leq b_k$	Presupuesto, equidad, capacidad, riesgo, estabilidad
Representación dinámica	Medidas ocupacionales d_π	Convierte el problema en LP/convexo interpretable
Interpretación económica	Dualidad y KKT (v, λ)	Precios sombra y restricciones activas auditables
Capa causal	g-recursión, supuestos secuenciales, agregación ϕ	Identifica efectos bajo intervención en datos observacionales
Evaluación sin despliegue	IS/DR, clipping, cotas OPE	Compara políticas candidatas con datos históricos
Control de error de modelo	Cotas tipo simulation lemma y suboptimalidad	Traduce error estadístico en riesgo de decisión
Robustez	$\mathcal{P}_{s,a}$, Bellman robusto	Protege contra incertidumbre y desalineación modelo-entorno
Seguridad	Restricciones deterministas/probabilísticas vs π_b	Evita deterioros inaceptables y regula despliegue
Aprendizaje restringido	Primal-dual estocástico	Implementación iterativa con garantías estándar

Componente del sistema	Objeto matemático	Papel funcional en IA pública
Monitoreo y adaptación	Drift, actualización de $\mathcal{D}_n$ y $\hat{\Psi}_n$	Cierra el ciclo institucional y empírico

Fuente: Elaboración propia.

Una IA para políticas públicas puede formularse con rigor matemático como un problema de optimización dinámica restringida, causalmente identificada de manera explícita, robustificada frente a incertidumbre y sujeta a reglas de seguridad respecto a una política base. El marco desarrollado no sustituye la deliberación política ni la gobernanza institucional, pero sí proporciona una estructura formal que permite discutir con precisión qué está siendo optimizado, bajo qué supuestos, con qué garantías y con qué riesgos residuales. Esa precisión es la condición mínima para que una IA pública sea técnicamente seria y normativamente auditable.

10. Metodología

El presente bloque metodológico documenta un protocolo de verificación integral, reproducible y auditable para un conjunto de proposiciones matemáticas y experimentos numéricos sobre procesos de decisión de Markov (MDP), medidas ocupacionales descontadas, operadores de Bellman, mejora de políticas, evaluación fuera de política (off-policy evaluation, OPE) mediante *importance sampling*, robustez bajo incertidumbre de transición, condiciones de seguridad respecto a una política *baseline*, penalización entrópica por divergencia KL, y un problema estático de asignación con utilidad cóncava resuelto por condiciones KKT y validado simbólicamente y numéricamente. El diseño metodológico se implementa en Colab con NumPy y SymPy, incorporando verificadores TRUE/FALSE por cada afirmación evaluada, así como un verificador global que sintetiza el estado del conjunto de pruebas.

El código utilizado en este trabajo está disponible en Google Colab en el siguiente enlace: Notebook en Google Colab.

La lógica del procedimiento no se limita a computar valores finales, sino que busca contrastar identidades, desigualdades, teoremas y condiciones suficientes mediante

múltiples capas de control. En particular, se emplean: i) resolución exacta de sistemas lineales para funciones de valor; ii) enumeración exacta de trayectorias de horizonte finito para verificar la insesgadez de estimadores IS; iii) simulación Monte Carlo con criterio estadístico explícito (3σ) para diagnóstico de cercanía empírica; iv) diferenciación simbólica con SymPy para comprobar derivadas de funciones de utilidad; v) verificación de restricciones y condiciones KKT por evaluación directa de igualdades y desigualdades; y vi) chequeos de contracción de operadores estándar y robustos mediante norma infinito.

El marco dinámico se basa en un MDP finito con espacio de estados $\mathcal{S}$ de cardinalidad $S = 3$, espacio de acciones $\mathcal{A}$ de cardinalidad $A = 2$, factor de descuento $\gamma = 0,9$ y distribución inicial $\mu \in \Delta(\mathcal{S})$. Las probabilidades de transición se representan como un tensor $P(a, s, s')$ y la recompensa instantánea como una matriz $r(s, a)$. Se consideran además tres políticas estocásticas: una política de comportamiento $\beta(a | s)$, una política objetivo $\pi(a | s)$ y una política alternativa $\pi_2(a | s)$. La estructura de datos se implementa con arreglos NumPy de dimensión fija, lo cual permite preservar transparencia en la trazabilidad y minimizar ambigüedades de indexación.

La función de transición inducida por una política π se construye como la matriz

$$P^\pi(s, s') = \sum_{a \in \mathcal{A}} \pi(a | s) P(a, s, s').$$

A partir de esta matriz, la función de valor $V^\pi \in \mathbb{R}^S$ se obtiene resolviendo el sistema lineal de Bellman

$$V^\pi = r^\pi + \gamma P^\pi V^\pi, \quad r^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a | s) r(s, a),$$

equivalentemente,

$$V^\pi = (I - \gamma P^\pi)^{-1} r^\pi.$$

La metodología privilegia esta resolución matricial exacta (hasta error numérico de máquina) en lugar de iteración de valor, con el fin de aislar los errores de verificación de los errores de convergencia iterativa. Esta decisión es especialmente útil cuando se pretende validar identidades analíticas como la equivalencia entre formulaciones basadas en Bellman y formulaciones basadas en medidas ocupacionales.

La medida ocupacional descontada normalizada se define, en su versión de estado-

acción, como

$$d_{\pi}(s, a) = x_{\pi}(s) \pi(a | s),$$

donde x_{π} es la distribución ocupacional por estado inducida por π ,

$$x_{\pi}^{\top} = (1 - \gamma) \mu^{\top} (I - \gamma P^{\pi})^{-1}.$$

Bajo esta normalización, $\sum_{s,a} d_{\pi}(s, a) = 1$, lo que facilita interpretar d_{π} como una distribución sobre pares estado-acción bajo descuento geométrico. La metodología incluye tres verificaciones esenciales asociadas a este objeto: no negatividad, suma unitaria y ecuación de flujo ocupacional. Esta última se expresa, para cada estado s , como

$$\sum_a d_{\pi}(s, a) = (1 - \gamma) \mu(s) + \gamma \sum_{s', a'} d_{\pi}(s', a') P(a', s', s).$$

Estas tres comprobaciones son estructuralmente importantes porque descartan errores frecuentes de implementación (por ejemplo, confusión entre filas y columnas, o normalizaciones incorrectas del vector ocupacional).

La identidad de valor basada en la medida ocupacional se verifica mediante

$$J(\pi) = \mu^{\top} V^{\pi} = \frac{1}{1 - \gamma} \sum_{s,a} d_{\pi}(s, a) r(s, a).$$

La estrategia metodológica consiste en computar $J(\pi)$ por ambas rutas de forma independiente y contrastar la diferencia absoluta contra una tolerancia numérica estricta. El hecho de que ambas cantidades coincidan con error numérico prácticamente nulo funciona como control cruzado de: i) la implementación del sistema de Bellman; ii) la construcción de d_{π} ; iii) el manejo del factor $(1 - \gamma)$ en la normalización.

El operador de Bellman para una política fija π se define como

$$(T^{\pi}V)(s) = \sum_a \pi(a | s) \left[r(s, a) + \gamma \sum_{s'} P(a, s, s') V(s') \right].$$

La metodología verifica numéricamente su propiedad de contracción en norma infinito:

$$\|T^{\pi}V_1 - T^{\pi}V_2\|_{\infty} \leq \gamma \|V_1 - V_2\|_{\infty},$$

utilizando dos vectores arbitrarios V_1, V_2 . Aunque un solo ejemplo no constituye

prueba general, sí constituye una validación de consistencia del operador implementado y de la manipulación de índices. Adicionalmente, se verifica que V^π satisface la ecuación de punto fijo $\|T^\pi V^\pi - V^\pi\|_\infty \approx 0$, lo cual refuerza la coherencia interna del bloque dinámico.

Para el teorema de mejora de política (*policy improvement theorem*), la metodología procede en cuatro pasos: i) computar V^π ; ii) obtener la función de acción-valor

$$Q^\pi(s, a) = r(s, a) + \gamma \sum_{s'} P(a, s, s') V^\pi(s');$$

iii) construir una política voraz π' que maximiza $Q^\pi(s, a)$ estado por estado; y iv) contrastar tanto la condición punto a punto

$$\mathbb{E}_{a \sim \pi'(\cdot|s)}[Q^\pi(s, a)] \geq V^\pi(s),$$

como la conclusión

$$V^{\pi'}(s) \geq V^\pi(s) \quad \forall s \in \mathcal{S}.$$

La elección de una política determinista voraz hace más nítida la interpretación del resultado, ya que la mejora puede observarse directamente en los márgenes por estado. El protocolo registra el margen mínimo de mejora en cada verificación, lo que permite cuantificar cuán lejos se está del umbral de igualdad.

La metodología también incluye la verificación numérica exacta del *performance difference lemma* (PDL), formulado como

$$J(\pi_2) - J(\pi) = \frac{1}{1 - \gamma} \sum_{s,a} d_{\pi_2}(s, a) A^\pi(s, a),$$

donde

$$A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$$

es la ventaja de π . Este bloque es especialmente relevante porque conecta evaluación de políticas, medidas ocupacionales y ventajas bajo una identidad exacta. La implementación calcula ambos lados de la igualdad de forma independiente y reporta la diferencia numérica. En términos metodológicos, este paso opera como una bisagra entre los resultados de Bellman y los análisis posteriores de seguridad y mejora respecto a *baseline*.

Con el objetivo de emular restricciones de política pública, se incorporan matrices de costo, uso de capacidad y cobertura mínima, denotadas genéricamente por $g(s, a)$, $h(s, a)$ y $c(s, a)$. A partir de d_π , se evalúan funcionales lineales descontados de la forma

$$G_g(\pi) = \frac{1}{1-\gamma} \sum_{s,a} d_\pi(s, a) g(s, a), \quad G_h(\pi) = \frac{1}{1-\gamma} \sum_{s,a} d_\pi(s, a) h(s, a),$$

$$G_c(\pi) = \frac{1}{1-\gamma} \sum_{s,a} d_\pi(s, a) c(s, a).$$

Las restricciones se expresan como

$$G_g(\pi) \leq B, \quad G_h(\pi) \leq \bar{H}, \quad G_c(\pi) \geq C_{\min}.$$

La metodología verifica cada una de estas desigualdades y reporta el margen de holgura, lo cual permite interpretar no sólo la factibilidad, sino también la robustez del cumplimiento bajo pequeñas perturbaciones.

En la parte de OPE, se establece un horizonte finito $T = 4$ y se verifica la identidad de insesgadez de *importance sampling* por enumeración exacta de trayectorias. Sea G_T el retorno descontado de horizonte finito y W_T el peso de razones de probabilidad entre política objetivo y política de comportamiento. La identidad contrastada es

$$\mathbb{E}_\beta[W_T G_T] = \mathbb{E}_\pi[G_T].$$

La metodología implementa una recursión exhaustiva sobre trayectorias posibles (s_t, a_t, s_{t+1}) para obtener ambos términos con precisión numérica alta. Esta decisión evita atribuir la validación de insesgadez a fluctuaciones Monte Carlo, separando una demostración numérica exacta de un diagnóstico estadístico posterior.

Tras la enumeración exacta, se realiza un diagnóstico Monte Carlo con $N_{mc} = 120,000$ trayectorias simuladas bajo β . Se computa la variable aleatoria $Y = W_T G_T$, el estimador $\hat{J}_{IS} = \frac{1}{N} \sum_i Y_i$, su desviación estándar muestral $\hat{\sigma}_Y$, el error estándar $\widehat{SE}(\hat{J}_{IS}) = \hat{\sigma}_Y / \sqrt{N}$, y se evalúa el criterio

$$\left| \hat{J}_{IS} - \mathbb{E}_\pi[G_T] \right| \leq 3 \widehat{SE}(\hat{J}_{IS}).$$

El verificador asociado no pretende demostrar inesgadez (ya demostrada por enumeración exacta), sino diagnosticar si la realización Monte Carlo observada es estadísticamente compatible con el valor exacto bajo una banda razonable. Esta distinción metodológica es central para evitar malinterpretar fluctuaciones muestrales como fallas teóricas.

En el bloque de robustez de modelo, se construyen una recompensa perturbada $\hat{r}$ y un tensor de transición perturbado $\hat{P}$, imponiendo renormalización de probabilidades para preservar validez estocástica. Se comparan las funciones de valor V^π y $\hat{V}^\pi$ bajo la misma política π , y se verifica una cota tipo *simulation lemma*:

$$\|\hat{V}^\pi - V^\pi\|_\infty \leq \frac{\varepsilon_r}{1 - \gamma} + \frac{2\gamma R_{\max}}{(1 - \gamma)^2} \varepsilon_P,$$

donde

$$\varepsilon_r = \|\hat{r} - r\|_\infty, \quad \varepsilon_P = \max_{s,a} \text{TV}(\hat{P}(\cdot | s, a), P(\cdot | s, a)), \quad R_{\max} = \max_{s,a} |r(s, a)|.$$

La distancia TV se calcula como variación total

$$\text{TV}(p, q) = \frac{1}{2} \|p - q\|_1.$$

El uso de una cota explícita permite vincular desviaciones paramétricas locales con errores globales en valor, proporcionando una lectura metodológica de sensibilidad de planificación bajo incertidumbre de modelo.

La verificación de contracción robusta emplea un operador de Bellman robusto de tipo *worst-case* sobre conjuntos de ambigüedad definidos por radios de variación total. Para cada par (s, a) , se considera un radio $\delta_{s,a}$ y se aproxima el problema interno de peor esperanza mediante una grilla sobre el simplex de dimensión tres:

$$\inf_{q \in \Delta(\mathcal{S}) : \text{TV}(q, P(\cdot | s, a)) \leq \delta_{s,a}} q^\top V.$$

Esta infimización se aproxima por búsqueda discreta sobre una grilla de resolución $m = 50$, lo que genera un conjunto finito de candidatos. El operador robusto se define entonces como

$$(T_R^\pi V)(s) = \sum_a \pi(a | s) \left[r(s, a) + \gamma \inf_{q : \text{TV}(q, P(\cdot | s, a)) \leq \delta_{s,a}} q^\top V \right].$$

La metodología compara $\|T_R^\pi V_1 - T_R^\pi V_2\|_\infty$ contra $\gamma\|V_1 - V_2\|_\infty$ para dos vectores de prueba. El carácter aproximado de la verificación se explicita, ya que el operador interno se discretiza por grilla, pero la prueba numérica sigue siendo informativa respecto a la consistencia del módulo robusto.

En el bloque de seguridad respecto a *baseline*, se fija $\pi_b = \beta$ como política de referencia. Se calculan $J(\pi_b)$ y $J(\pi_{\text{new}})$, donde π_{new} corresponde a la política voraz antes construida, y se contrasta la condición directa de seguridad con tolerancia $\varepsilon_{\text{safe}}$:

$$J(\pi_{\text{new}}) \geq J(\pi_b) - \varepsilon_{\text{safe}}.$$

Además, se evalúa una condición suficiente expresada mediante ventajas respecto a la *baseline*. Si

$$A^{\pi_b}(s, a) = Q^{\pi_b}(s, a) - V^{\pi_b}(s),$$

y $d_{\pi_{\text{new}}}$ denota la medida ocupacional de la política candidata, se verifica el término

$$\sum_{s,a} d_{\pi_{\text{new}}}(s, a) A^{\pi_b}(s, a),$$

comparándolo con el umbral $-(1 - \gamma)\varepsilon_{\text{safe}}$. Este doble enfoque (directo y suficiente) aporta una capa metodológica importante: la primera verificación usa el valor exacto, mientras la segunda usa una expresión estructuralmente interpretable para análisis de diseño de políticas seguras.

La penalización entrópica se estudia mediante la funcional

$$f(\pi) = -\tau \sum_s \nu(s) \text{KL}(\pi(\cdot | s) \| \pi_b(\cdot | s)),$$

con $\tau > 0$ y pesos $\nu \in \Delta(\mathcal{S})$. Se verifica numéricamente la concavidad vía una desigualdad de Jensen para dos políticas π_A, π_B y una mezcla $\pi_M = \lambda\pi_A + (1 - \lambda)\pi_B$:

$$f(\pi_M) \geq \lambda f(\pi_A) + (1 - \lambda) f(\pi_B).$$

La divergencia KL se computa fila por fila con recorte numérico (*clipping*) para evitar problemas de $\log 0$, seguido de renormalización. Esta práctica metodológica preserva estabilidad numérica sin alterar sustantivamente la lógica de la verificación en un contexto donde todas las probabilidades permanecen estrictamente positivas por construcción.

La parte estática introduce una función de utilidad cóncava

$$u(x) = a \log(1 + bx),$$

y utiliza SymPy para verificar simbólicamente las derivadas

$$u'(x) = \frac{ab}{1 + bx}, \quad u''(x) = -\frac{ab^2}{(1 + bx)^2}.$$

Estas igualdades constituyen una mini-validación formal del módulo simbólico y de la convexidad/concavidad requerida para el problema de asignación posterior. El uso combinado de cálculo simbólico y resolución numérica ofrece una metodología híbrida útil en contextos de diseño de algoritmos con fundamento analítico.

El problema de asignación con restricción presupuestaria se formula como

$$\max_{x_i \geq 0} \sum_{i=1}^n a_i \log(1 + b_i x_i) \quad \text{sujeto a} \quad \sum_{i=1}^n x_i = B.$$

La metodología implementa una solución cerrada tipo *water-filling* derivada de las KKT:

$$x_i^*(\lambda) = \max \left\{ 0, \frac{a_i}{\lambda} - \frac{1}{b_i} \right\},$$

con λ^* obtenido por bisección para satisfacer la restricción de presupuesto. Posteriormente se verifican: i) igualdad de marginales para componentes activos,

$$\frac{a_i b_i}{1 + b_i x_i^*} = \lambda^* \quad \text{si } x_i^* > 0;$$

ii) condición de no exceso para componentes inactivos,

$$\frac{a_i b_i}{1 + b_i x_i^*} \leq \lambda^* \quad \text{si } x_i^* = 0;$$

iii) cumplimiento vinculante del presupuesto; y iv) evidencia empírica de optimalidad comparando $U(x^*)$ con el mejor valor hallado en una búsqueda aleatoria sobre el simplex presupuestario. Este último paso no es una prueba de optimalidad global (ya garantizada por concavidad y KKT), pero sí una corroboración computacional adicional.

Se incorpora una cota de sensibilidad normativa bajo pesos alternativos. Dado un vector $F(\pi) \in \mathbb{R}^n$ y dos vectores de ponderación $w, \tilde{w} \in \Delta_n$, se evalúa la desigualdad

$$\left| w^\top F(\pi) - \tilde{w}^\top F(\pi) \right| \leq \|w - \tilde{w}\|_1 \|F(\pi)\|_\infty.$$

Metodológicamente, esta cota permite interpretar cómo cambios en ponderaciones normativas o prioridades distributivas pueden alterar una evaluación agregada sin necesidad de reestimar toda la estructura dinámica. Su inclusión amplía el alcance del protocolo desde consistencia matemática estricta hacia análisis de sensibilidad institucional.

La Tabla 9 resume la parametrización base del módulo dinámico. La Tabla 10 resume los bloques de verificación implementados y su propósito. Ambas tablas cumplen un rol metodológico de documentación reproducible, permitiendo reconstruir la lógica del experimento sin inspeccionar directamente el código.

Tabla 9: Parámetros y objetos del MDP base utilizados en las verificaciones

Elemento	Descripción metodológica
$S = 3, A = 2$	MDP finito de pequeña dimensión para posibilitar enumeración exacta en OPE y trazabilidad completa de resultados.
$\gamma = 0,9$	Factor de descuento alto, suficientemente representativo de problemas de planificación intertemporal con persistencia.
μ	Distribución inicial sobre estados usada tanto en Bellman como en medida ocupacional y en OPE.
$P(a, s, s')$	Tensor de transición con dos acciones y tres estados; todas las filas son estocásticas y estrictamente positivas.
$r(s, a)$	Matriz de recompensas instantáneas empleada en Bellman, Q^π , PDL y OPE.
β	Política de comportamiento para OPE y política <i>baseline</i> en el bloque de seguridad.
π	Política objetivo principal para verificación de medidas ocupacionales, Bellman, PDL y robustez.
π_2	Política comparativa para la verificación numérica exacta del <i>performance difference lemma</i> .

Fuente: Elaboración propia.

Desde el punto de vista de reproducibilidad, el protocolo se beneficia de tres elecciones metodológicas adicionales. En primer lugar, se fija una semilla explícita para los

Tabla 10: Bloques de verificación incluidos en el protocolo integral

Bloque	Objeto matemático principal	Tipo de verificación
Medidas ocupacionales	d_π , flujo, identidad de valor	Igualdades exactas numéricas con tolerancias
Bellman	Contracción y punto fijo	Desigualdad en norma ∞ y residuo de punto fijo
Mejora de política	Q^π , política voraz	Condición punto a punto y comparación de V
PDL	$J(\pi_2) - J(\pi)$ y ventajas	Identidad exacta numérica
Restricciones lineales	Presupuesto, capacidad, cobertura	Desigualdades con márgenes de holgura
OPE exacto	$\mathbb{E}_\beta[W_T G_T]$ vs $\mathbb{E}_\pi[G_T]$	Enumeración exacta de trayectorias
OPE Monte Carlo	$\hat{J}_{IS}$	Diagnóstico estadístico con regla 3σ
Robustez de modelo	Cota tipo <i>simulation lemma</i>	Verificación de cota superior
Bellman robusto	T_R^π	Contracción aproximada en grilla simplex
Seguridad	Comparación vs <i>baseline</i> y ventajas	Condición directa y condición suficiente
Penalización KL	Concavidad de $-\tau \mathbb{E}[\text{KL}]$	Jensen numérico
KKT + utilidad	Derivadas, KKT, presupuesto	SymPy + validación numérica + búsqueda aleatoria
Sensibilidad normativa	Cota ℓ_1 - ℓ_∞	Desigualdad exacta numérica

Fuente: Elaboración propia.

generadores pseudoaleatorios en los bloques Monte Carlo y de búsqueda aleatoria, lo que permite replicar exactamente la secuencia de resultados de una corrida dada. En segundo lugar, los objetos intermedios relevantes (por ejemplo, d_π , V^π , Q^π , A^π , márgenes de restricciones, estadísticas de OPE y multiplicadores KKT) se imprimen en consola con formato legible, facilitando auditoría paso a paso y diagnóstico en caso de futuras modificaciones. En tercer lugar, el sistema de verificadores TRUE/FALSE se diseña como una capa declarativa: cada chequeo registra nombre, valor booleano y detalle cuantitativo, y el verificador global resume el estado del conjunto. Esta arquitectura hace que el protocolo sea útil no sólo para análisis académico, sino también para *testing* de regresión en desarrollo de software científico.

Cabe enfatizar que varias verificaciones aquí presentadas son *demonstraciones numéricas* o *corroboraciones computacionales* de proposiciones teóricas, no demostraciones formales universales. La metodología, por tanto, debe leerse como una estrategia de validación computacional rigurosa en un entorno controlado. Su fuerza principal reside en la combinación de exactitud en casos pequeños (por enumeración o solución cerrada), robustez estadística en bloques de simulación, y consistencia transversal entre formulaciones equivalentes. Este diseño es particularmente pertinente en aplicaciones de IA para política pública, donde la transparencia, la trazabilidad y la verificabilidad son tan importantes como el valor predictivo u operativo.

11. Resultados

Los resultados obtenidos muestran una consistencia interna notable entre los distintos módulos del protocolo, culminando en un verificador global TRUE con 26/26 comprobaciones superadas. Este desenlace es relevante por dos razones. Primero, porque el conjunto de verificaciones abarca objetos heterogéneos (dinámicos, robustos, estadísticos, simbólicos y de optimización estática) y, por tanto, reduce la probabilidad de que el éxito se deba a una coincidencia local o a un error compensatorio. Segundo, porque varias de las comprobaciones contrastan identidades por rutas computacionales independientes, lo que fortalece la inferencia de corrección de implementación.

El bloque de medidas ocupacionales descontadas produce una distribución $d_\pi(s, a)$ estrictamente no negativa y correctamente normalizada, con suma total igual a uno dentro de precisión numérica de máquina. Esta propiedad no es trivial en términos

computacionales, dado que la construcción de d_π requiere una inversión matricial y una composición con la política. Además, la ecuación de flujo ocupacional se verifica con error máximo de orden 10^{-16} , lo que sugiere ausencia de errores de indexación o transposición en la manipulación del tensor de transición. La identidad de valor basada en ocupación también coincide exactamente con el valor obtenido vía Bellman, con diferencia prácticamente nula. En conjunto, estos resultados indican que la representación ocupacional y la representación de valor están plenamente alineadas en la implementación.

La Tabla 11 resume los principales resultados numéricos del bloque ocupacional y del operador de Bellman. Se observa que la función de valor V^π presenta niveles relativamente altos en los tres estados, consistentes con un factor de descuento elevado ($\gamma = 0,9$) y recompensas positivas predominantes. La mayor valoración del estado 2 frente al estado 1 en V^π es coherente con la combinación de recompensas por acción y estructura de transición bajo la política π , aunque el orden exacto entre estados refleja la interacción no lineal entre ambas.

Tabla 11: Resultados del bloque ocupacional y Bellman para la política π

Magnitud	Valor reportado	Interpretación
$\sum_{s,a} d_\pi(s, a)$	1,000000000000	Normalización correcta de la medida ocupacional descontada.
Error de flujo ocupacional ($\max LHS - RHS $)	$1,110 \times 10^{-16}$	Consistencia numérica esencialmente exacta.
V^π	[8,723433, 9,450659, 9,035601]	Valor esperado descontado por estado bajo π .
$J(\pi)$ vía Bellman	9,004034297510	Valor agregado bajo distribución inicial μ .
$J(\pi)$ vía ocupación	9,004034297510	Coincide con Bellman; válida identidad de valor.

Fuente: Elaboración propia.

En el bloque de Bellman, la verificación de contracción se satisface con holgura. El valor observado de $\|T^\pi V_1 - T^\pi V_2\|_\infty$ es significativamente menor que $\gamma \|V_1 - V_2\|_\infty$, lo que no sólo confirma la desigualdad en el ejemplo ensayado, sino que además sugiere una contracción efectiva relativamente fuerte para los vectores seleccionados. El residuo de punto fijo $\|T^\pi(V^\pi) - V^\pi\|_\infty$ es de orden 10^{-15} , consistente con errores

de redondeo de doble precisión. Este resultado respalda la corrección del solucionador matricial y de la implementación del operador T^π .

La construcción de la política voraz a partir de Q^π produce una mejora clara. La condición $\mathbb{E}_{\pi'}[Q^\pi(\cdot, \cdot) | s] \geq V^\pi(s)$ se verifica en los tres estados con margen mínimo positivo, y la comparación de funciones de valor muestra incrementos sustanciales, con un margen mínimo de mejora superior a 1,95 unidades. La política voraz resultante es determinista y selecciona la acción 0 en el estado 1, la acción 1 en el estado 2 y la acción 1 en el estado 3 (siguiendo la indexación base 0, 1, 2 usada en código, esto corresponde al patrón $[0, 1, 1]$). La magnitud de la mejora sugiere que la política π inicial estaba lejos de ser óptima en el MDP considerado.

La Tabla 12 sintetiza los resultados de mejora de política y del *performance difference lemma*. Es especialmente importante la coincidencia prácticamente exacta entre ambos lados del PDL, ya que esta identidad enlaza el análisis local (ventajas A^π) con el cambio global de desempeño $J(\pi_2) - J(\pi)$. En la práctica, esta verificación provee una garantía fuerte de consistencia en el cálculo de Q^π , V^π , A^π , d_{π_2} y $J(\cdot)$.

Tabla 12: Resultados de mejora de política y *performance difference lemma*

Magnitud	Valor reportado	Lectura
Margen mínimo de condición de mejora	$1,431 \times 10^{-1}$	La política voraz mejora Q^π punto a punto.
Margen mínimo en $V^{\pi'} - V^\pi$	1,951	Mejora sustancial en valor por estado.
$V^{\pi'}$	[10,986301, 11,534247, 10,986301]	Política voraz fuertemente superior a π .
$J(\pi_2) - J(\pi)$	0,696260318843	Diferencia global de desempeño.
$\text{PDL RHS} = \frac{1}{1-\gamma} \sum d_{\pi_2} A^\pi$	0,696260318843	Coincidencia exacta con LHS (error $\sim 10^{-15}$).

Fuente: Elaboración propia.

En el bloque de restricciones lineales, la política π resulta factible respecto a las tres restricciones propuestas. El costo agregado descontado G_{cost} queda por debajo del presupuesto B con una holgura de aproximadamente 0,077, la restricción de capacidad presenta una holgura más amplia ($\approx 0,570$), y la cobertura mínima supera el umbral en aproximadamente 1,761. Estas magnitudes son metodológicamente útiles porque permiten inferir que, en este ejemplo, la restricción más ajustada es el presupuesto.

En aplicaciones reales, esta lectura puede orientar qué restricción es más probable que se active tras una mejora de política o una perturbación de modelo.

La factibilidad simultánea de las restricciones también demuestra la utilidad de las medidas ocupacionales como objeto unificador: una vez construido d_π , múltiples métricas de política se evalúan como funcionales lineales sobre el mismo soporte estado-acción. Esto reduce complejidad computacional y facilita el análisis de sensibilidad, pues la introducción de nuevas métricas $m(s, a)$ no exige re-simular trayectorias ni re-resolver Bellman, sino sólo computar $\frac{1}{1-\gamma} \sum d_\pi m$.

Los resultados de OPE son particularmente ilustrativos. La identidad exacta de insesgadez por enumeración se satisface con diferencia nula (dentro de precisión numérica), mostrando que

$$\mathbb{E}_\beta[W_T G_T] = \mathbb{E}_\pi[G_T]$$

para el horizonte $T = 4$ considerado. Este resultado tiene dos implicaciones. Primero, confirma la implementación correcta de la recursión de enumeración y de los pesos IS. Segundo, valida la consistencia de la definición de retorno descontado y su alineación temporal con los pesos.

En el diagnóstico Monte Carlo, el estimador IS $\hat{J}_{IS}$ difiere del valor exacto en aproximadamente $1,96 \times 10^{-2}$, mientras que la banda de $3\widehat{SE}$ es aproximadamente $9,82 \times 10^{-2}$. Dado que la diferencia observada es bastante menor que la banda, el diagnóstico se clasifica como satisfactorio. La desviación estándar de $Y = W_T G_T$ es relativamente alta en comparación con la escala del retorno, lo que evidencia la variabilidad inherente del estimador IS incluso en un problema pequeño. Asimismo, el máximo observado de los pesos W (del orden de 10^2) ilustra, en miniatura, el fenómeno bien conocido de colas pesadas o alta varianza potencial en OPE por razones de probabilidad. El hecho de que aun así el diagnóstico pase con $N = 120,000$ muestra que el tamaño muestral elegido fue suficiente para estabilizar la estimación en este caso.

La Tabla 13 resume los resultados exactos y Monte Carlo del bloque OPE. La lectura conjunta de esta tabla sugiere una lección metodológica importante: cuando es posible disponer de un valor exacto (por enumeración en problemas pequeños o por casos sintéticos), conviene usarlo como referencia para calibrar criterios de diagnóstico Monte Carlo antes de escalar a problemas donde dicha referencia no existe.

En el bloque de robustez por mala especificación del modelo, la cota tipo *simulation*

Tabla 13: Resultados de OPE: identidad exacta y diagnóstico Monte Carlo

Magnitud	Valor reportado	Interpretación
Horizonte T	4	OPE en horizonte finito con descuento.
$\mathbb{E}_\beta[W_T G_T]$ exacto	3,618135325759	Lado IS exacto por enumeración.
$\mathbb{E}_\pi[G_T]$ exacto	3,618135325759	Lado objetivo exacto por enumeración.
$\hat{J}_{IS}$ Monte Carlo	3,598502169240	Estimador IS sobre 120,000 trayectorias.
Diferencia absoluta vs exacto	0,01963316	Error muestral observado.
$3\widehat{SE}(\hat{J}_{IS})$	0,098187	Banda de diagnóstico estadístico.
Resultado del diagnóstico	TRUE	La diferencia cae dentro de la banda 3σ .

Fuente: Elaboración propia.

lemma se verifica con margen amplio. La diferencia real $\|\hat{V}^\pi - V^\pi\|_\infty$ es aproximadamente 0,116, mientras que la cota superior calculada alcanza 5,44. Esta holgura grande era esperable: las cotas uniformes en norma infinito suelen ser conservadoras, especialmente cuando agregan errores de recompensa y transición de manera worst-case. El resultado es, sin embargo, metodológicamente valioso porque confirma que la implementación de la cota y de las distancias de variación total es consistente. Además, evidencia una situación clásica en análisis robusto: una cota puede ser válida y útil para garantías, aun cuando no sea ajustada.

La perturbación de transiciones se diseñó de modo que todas las distancias de variación total por par (s, a) fueran exactamente 0,02. Esto simplifica la interpretación del parámetro ε_P y permite atribuir el tamaño de la cota a la estructura analítica del término $\frac{2\gamma R_{\max}}{(1-\gamma)^2}\varepsilon_P$, que crece rápidamente al acercarse γ a 1. En términos prácticos, el resultado ilustra que en problemas con horizontes efectivos largos (alto γ), pequeñas perturbaciones de transición pueden inducir garantías teóricas muy conservadoras.

La verificación de contracción del operador robusto T_R^π , aproximado por grilla sobre el simplex, también se satisface con holgura. El valor observado de $\|T_R^\pi(V_1) - T_R^\pi(V_2)\|_\infty$ es aproximadamente 0,8442, frente a un límite $\gamma\|V_1 - V_2\|_\infty = 1,8$. El margen negativo amplio ($lhs - rhs < 0$) sugiere que la contracción efectiva del operador robusto

en el ejemplo analizado no es cercana al límite teórico. No obstante, al tratarse de una verificación aproximada por grilla, la interpretación debe mantenerse prudente: el resultado corrobora la consistencia numérica del módulo robusto, pero no sustituye una demostración analítica general del operador bajo la clase de conjuntos de ambigüedad utilizada.

En seguridad respecto a *baseline*, los resultados son muy favorables para la política nueva (voraz). El valor agregado $J(\pi_{\text{new}})$ supera ampliamente a $J(\pi_b) - \varepsilon_{\text{safe}}$, con un margen cercano a 2,97, muy superior al umbral de tolerancia $\varepsilon_{\text{safe}} = 0,05$. La condición suficiente basada en ventaja también se cumple con holgura, ya que el término $\sum d_{\pi_{\text{new}}} A^{\pi_b}$ es positivo y claramente mayor que el umbral negativo permitido. Esta coincidencia entre condición directa y condición suficiente tiene implicancias metodológicas útiles: cuando la evaluación exacta de $J(\pi_{\text{new}})$ sea costosa o imposible, las condiciones de ventaja pueden funcionar como *screening* para filtrar políticas seguras candidatas.

La Tabla 14 agrupa los resultados de robustez y seguridad. Su lectura conjunta sugiere que el protocolo no sólo es capaz de verificar consistencia matemática, sino también de capturar dimensiones de interés aplicado: sensibilidad al modelo y seguridad de despliegue respecto a una política de referencia.

Tabla 14: Resultados de robustez de modelo, Bellman robusto y seguridad respecto a *baseline*

Magnitud	Valor reportado	Lectura
$\ \hat{V}^{\pi} - V^{\pi}\ _{\infty}$	0,116184	Error real bajo perturbación de r y P .
Cota <i>simulation lemma</i>	5,440000	Garantía válida pero conservadora.
$\ T_R^{\pi}(V_1) - T_R^{\pi}(V_2)\ _{\infty}$	0,844200	LHS de contracción robusta aproximada.
$\gamma\ V_1 - V_2\ _{\infty}$	1,800000	RHS de contracción robusta.
$J(\pi_b)$	8,231571768763	Valor de la política <i>baseline</i> .
$J(\pi_{\text{new}})$	11,150684931507	Valor de la política candidata voraz.
Margen de seguridad directo	2,969113	Cumplimiento muy holgado de la condición directa.
Margen de condición suficiente por ventaja	0,296911	Cumplimiento holgado de la condición estructural.

Fuente: Elaboración propia.

La verificación de concavidad de la penalización $-\tau \mathbb{E}_v[\text{KL}(\pi \| \pi_b)]$ mediante Jensen produce una brecha positiva (*gap*) de aproximadamente $2,997 \times 10^{-2}$, lo que confirma empíricamente la concavidad en el ejemplo analizado. Este resultado es conceptualmente importante porque justifica el uso de términos de regularización entrópica con buena estructura de optimización en diseños de políticas. En particular, cuando tales penalizaciones se incorporan a objetivos compuestos, la concavidad de esta parte del funcional puede contribuir a estabilidad y tractabilidad de algoritmos de actualización.

En el bloque simbólico y de KKT, SymPy verifica exactamente las derivadas primera y segunda de $u(x) = a \log(1 + bx)$. Esta parte, aunque breve, funciona como una validación del canal simbólico y como respaldo formal de la concavidad del problema estático posterior. Luego, la solución numérica del problema de asignación presupuestaria produce un multiplicador $\lambda^* \approx 0,65483$ y una asignación x^* con todos los componentes activos. El hecho de que todos los componentes resulten activos implica que el régimen KKT observado es interior, por lo que la condición de igualdad de marginales se verifica en los cuatro componentes. La coincidencia de marginales con λ^* hasta error de máquina es una evidencia fuerte de correcta implementación de la solución tipo *water-filling* y del algoritmo de bisección.

La restricción presupuestaria se satisface como igualdad con error de orden 10^{-16} , lo que confirma el éxito del método de bisección en encontrar λ^* . Adicionalmente, la búsqueda aleatoria sobre 30,000 puntos del simplex presupuestario no encuentra una solución mejor que x^* , con una diferencia de utilidad a favor de x^* de aproximadamente $1,98 \times 10^{-4}$. Aunque esta evidencia aleatoria no es necesaria para concluir optimalidad (ya garantizada por concavidad y KKT), sí aporta una comprobación empírica intuitiva y útil para auditoría externa.

La cota de sensibilidad normativa se verifica con margen claro. La diferencia absoluta entre agregaciones $w^\top F(\pi)$ y $\tilde{w}^\top F(\pi)$ es 0,4, mientras que la cota $\|w - \tilde{w}\|_1 \|F(\pi)\|_\infty$ es 0,84. Este resultado ilustra que la desigualdad ℓ_1 - ℓ_∞ opera como garantía general y, al igual que otras cotas uniformes, puede no ser ajustada. La interpretación aplicada es directa: cambios en ponderaciones normativas pueden alterar la evaluación agregada, pero su efecto máximo está acotado por una combinación simple de distancia entre pesos y amplitud del vector de resultados.

La Tabla 15 resume los resultados del bloque estático y de sensibilidad normativa.

En conjunto con los bloques dinámicos previos, estos resultados muestran que el protocolo logra integrar validación de teoría dinámica de decisiones, estadística de evaluación fuera de política, análisis robusto, regularización entrópica y optimización estática bajo una misma arquitectura de verificación.

Tabla 15: Resultados de penalización KL, derivación simbólica, KKT y sensibilidad normativa

Magnitud	Valor reportado	Interpretación
Gap de Jensen para $-\tau\mathbb{E}[\text{KL}]$	$2,997 \times 10^{-2}$	Concavidad confirmada numéricamente en el ejemplo.
Verificación simbólica de $u'(x)$ y $u''(x)$	TRUE, TRUE	Derivadas exactas obtenidas con SymPy.
λ^* (KKT)	0,654831508451	Multiplicador óptimo del presupuesto.
Cumplimiento de igualdad de marginales activos	Error $\sim 10^{-16}$	KKT satisfecha con precisión de máquina.
Presupuesto en x^*	$\sum_i x_i^* = 4,2$	Restricción vinculante correctamente satisfecha.
Evidencia aleatoria vs x^*	$U(x^*)$ $U(z_{\text{máx}})$ $1,98 \times 10^{-4}$	— $\approx$ No se encontró mejora aleatoria sobre x^* .
Cota de sensibilidad normativa	$0,4 \leq 0,84$	Desigualdad ℓ_1 - ℓ_∞ verificada.

Fuente: Elaboración propia.

Una manera sintética de representar el núcleo del resultado global es mediante el vector de verificaciones

$$\mathbf{c} = (c_1, \dots, c_{26}), \quad c_i \in \{0, 1\},$$

donde $c_i = 1$ si la verificación i -ésima se satisface y 0 en caso contrario. En la corrida reportada se obtuvo

$$\sum_{i=1}^{26} c_i = 26,$$

por lo que el verificador global toma el valor

$$\mathcal{V}_{\text{global}} = \mathbf{1} \left\{ \sum_{i=1}^{26} c_i = 26 \right\} = 1.$$

Esta formulación, aunque simple, es útil para estandarizar comparaciones entre versiones del código o entre escenarios de parametrización distintos. En un esquema de desarrollo iterativo, una degradación de $\mathcal{V}_{\text{global}}$ de 1 a 0 puede interpretarse como una alerta automática de regresión.

La Tabla 16 presenta un resumen consolidado de los 26 chequeos, con una descripción breve y su estado. Esta tabla puede funcionar como apéndice ejecutivo dentro de un documento técnico, permitiendo a un lector identificar de inmediato qué propiedades se verificaron sin recorrer toda la salida de consola.

Tabla 16: Resumen consolidado de las 26 verificaciones y su estado

#	Verificación	Estado
1	No negatividad de d_{π}	TRUE
2	Suma unitaria de d_{π}	TRUE
3	Ecuación de flujo ocupacional	TRUE
4	Identidad de valor con ocupación	TRUE
5	Contracción de T^{π} (ejemplo numérico)	TRUE
6	Punto fijo de Bellman V^{π}	TRUE
7	Condición de mejora $\mathbb{E}_{\pi'} Q^{\pi} \geq V^{\pi}$	TRUE
8	Conclusión $V^{\pi'} \geq V^{\pi}$	TRUE
9	<i>Performance difference lemma</i> exacto	TRUE
10	Restricción de presupuesto (ejemplo)	TRUE
11	Restricción de capacidad (ejemplo)	TRUE
12	Restricción de cobertura mínima (ejemplo)	TRUE
13	Inssegadez IS por enumeración exacta	TRUE
14	Diagnóstico Monte Carlo IS cercano al exacto	TRUE
15	Cota tipo <i>simulation lemma</i> (política fija)	TRUE
16	Contracción robusta (aprox. grilla simplex)	TRUE
17	Seguridad directa vs <i>baseline</i>	TRUE
18	Condición suficiente de seguridad por ventaja	TRUE
19	Concavidad de $-\tau \mathbb{E}[\text{KL}]$ (Jensen numérico)	TRUE
20	Verificación simbólica de $u'(x)$	TRUE
21	Verificación simbólica de $u''(x)$	TRUE
22	KKT: igualdad en componentes activos	TRUE
23	KKT: no exceso en componentes inactivos	TRUE
24	Presupuesto vinculante en x^*	TRUE
25	No mejora aleatoria sobre x^* (evidencia)	TRUE
26	Cota ℓ_1 - ℓ_{∞} de sensibilidad normativa	TRUE
Verificador global $\mathcal{V}_{\text{global}}$		TRUE (26/26)

Fuente: Elaboración propia.

Desde una perspectiva interpretativa, los resultados sugieren que el entorno de prueba construido es adecuado como banco de validación para desarrollos posteriores más ambiciosos. La coexistencia de identidades exactas (por ejemplo, medida ocupacional, PDL, OPE por enumeración), desigualdades teóricas (contracción, cotas de robustez, sensibilidad normativa), criterios estadísticos (diagnóstico Monte Carlo) y condiciones de optimalidad (KKT) en un mismo script favorece una cultura de verificación integral que suele faltar en implementaciones aplicadas de IA para decisiones públicas. En particular, la explicitación de márgenes de holgura y residuos numéricos aporta información más rica que un simple estado binario.

Al mismo tiempo, los resultados invitan a distinguir cuidadosamente entre distintos niveles de evidencia. Algunas comprobaciones son esencialmente exactas dentro de aritmética de punto flotante (por ejemplo, PDL, flujo ocupacional, derivadas simbólicas), mientras que otras son diagnósticas o aproximadas (por ejemplo, Monte Carlo IS y contracción robusta por grilla). Esta heterogeneidad no debilita el protocolo; por el contrario, lo hace más honesto metodológicamente, siempre que se explicita el estatus de cada prueba, como se ha hecho aquí.

12. Conclusión

El desarrollo de una inteligencia artificial para la optimización de políticas públicas exige un marco mucho más exigente que el habitual en problemas de predicción o control en entornos cerrados. A lo largo de este trabajo se ha mostrado que una formulación rigurosa debe integrar, de manera simultánea, dinámica secuencial, restricciones normativas, identificación causal, evaluación contrafactual, robustez frente a incertidumbre y seguridad respecto a políticas base. La idea central que emerge es que una IA pública no puede ser entendida como un mero algoritmo de maximización, sino como una arquitectura formal compuesta por capas distintas pero interdependientes: una capa causal-estadística que identifica y estima el problema, una capa de optimización que calcula políticas factibles y una capa normativa-institucional que fija los límites, prioridades y condiciones de despliegue.

Desde el punto de vista matemático, el documento ha establecido que los procesos de decisión de Markov restringidos, su representación mediante medidas ocupacionales y su formulación primal-dual proporcionan una base especialmente poderosa

para modelar políticas públicas dinámicas. Esta representación permite probar existencia de soluciones óptimas, realizabilidad de medidas ocupacionales, dualidad fuerte, condiciones KKT e interpretación de multiplicadores como precios sombra de restricciones públicas. Asimismo, el uso del lema de diferencia de desempeño y de resultados de mejora de política permite conectar el análisis de optimalidad con garantías de no deterioro respecto a un régimen base.

No obstante, también se ha enfatizado que la corrección interna del optimizador no basta. La validez de la política recomendada depende críticamente de la identificación causal de transiciones, recompensas y restricciones a partir de datos históricos. Sin esa capa causal, la IA puede optimizar patrones observados que no representan efectos de intervención genuinos. De ahí la importancia de la g-recursión, la ignorabilidad secuencial, la positividad y la suficiencia de la agregación al pasar desde historias completas a estados Markovianos agregados. Del mismo modo, la evaluación fuera de política, la doble robustez y las cotas de error por mala especificación del modelo son esenciales para cuantificar el riesgo de optimizar sobre representaciones imperfectas del sistema social.

El análisis también ha mostrado que la robustez no es una característica opcional, sino una exigencia constitutiva del uso público de la IA. La incertidumbre sobre transiciones, los shocks exógenos, la no estacionariedad y los límites de transportabilidad hacen necesario trabajar con conjuntos robustos, reglas de seguridad probabilística y mecanismos de regularización respecto a políticas base. A ello se suma la necesidad de incorporar explícitamente equidad, implementabilidad y estabilidad institucional, porque una política puede ser eficiente en términos agregados y, sin embargo, distributivamente regresiva, administrativamente inviable o políticamente desestabilizadora. La matemática permite formalizar estos conflictos, detectar incompatibilidades de factibilidad y hacer explícitos los costos marginales de endurecer o relajar restricciones.

En términos sustantivos, el resultado general del trabajo es una teoría de la IA pública como operador auditable de decisión bajo restricciones. Esta teoría no promete eliminar el juicio político ni sustituir la deliberación democrática. Por el contrario, muestra que incluso el sistema algorítmico más sofisticado necesita parámetros normativos explícitos, validación causal, reglas de seguridad y mecanismos de supervisión humana. La inteligencia artificial puede fortalecer la racionalidad pública sólo si sus supuestos, objetivos, restricciones y márgenes de error son transparentes y revisables.

En consecuencia, la principal conclusión es doble. Primero, sí es posible construir una base matemática para sistemas de IA orientados a la optimización de políticas públicas. Segundo, esa base sólo es legítima cuando distingue con claridad entre predicción y causalidad, entre optimalidad técnica y aceptabilidad normativa, y entre eficiencia computacional y responsabilidad institucional. La verdadera promesa de la IA en el sector público no reside en automatizar la política, sino en hacer explícitas, trazables y formalmente evaluables las condiciones bajo las cuales una decisión pública puede considerarse eficiente, segura, robusta y socialmente justificable.

Referencias

- Athey, S., & Wager, S. (2021). Policy Learning With Observational Data. *Econometrica*, 89(1), 133–161.
- Buckley, M., Papathanasiou, K., & Spanopoulos, A. (2025). *Primal-Dual Sample Complexity Bounds for Constrained Markov Decision Processes with Multiple Constraints*. arXiv preprint arXiv:2503.06751v1.
- Chen, K., Wei, H., Deng, Z., & Lin, S. (2026). *Towards Fast Safe Online Reinforcement Learning via Policy Finetuning*. Transactions on Machine Learning Research. arXiv preprint arXiv:2412.04426v3.
- De Moor, J., Weytjens, H., De Smedt, J., & De Weerd, J. (2026). *SCOPE: Sequential Causal Optimization of Process Interventions*. arXiv preprint arXiv:2512.17629v3.
- Ertefaie, A., & Strawderman, R. L. (2018). Constructing dynamic treatment regimes over indefinite time horizons. *Biometrika*, 105(4), 963–977.
- Feinberg, E. A., Jaśkiewicz, A., & Nowak, A. S. (2019). *Constrained discounted Markov decision processes with Borel state spaces*. arXiv preprint arXiv:1806.00190v7.
- Kalagarla, K. C., Jain, R., & Nuzzo, P. (2021). *A Sample-Efficient Algorithm for Episodic Finite-Horizon MDP with Constraints*. The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21).
- Sawarni, A., Jin, J., Whitehouse, J., & Syrgkanis, V. (2026). *Policy Learning with Abstention*. arXiv preprint arXiv:2510.19672v3.
- Stradi, F. E., Lunghi, A., Castiglioni, M., Marchesi, A., & Gatti, N. (2024). *Learning Constrained Markov Decision Processes With Non-stationary Rewards and Constraints*. arXiv preprint arXiv:2405.14372v2.
- Thomas, P. S., Theodorou, G., & Ghavamzadeh, M. (2015). High Confidence Off-Policy

- Evaluation. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Wendl, M., As, Y., Prajapat, M., Pollak, A., Coros, S., & Krause, A. (2026). *Safe Exploration via Policy Priors*. Published as a conference paper at ICLR 2026. arXiv preprint arXiv:2601.19612v1.
- Xiao, X., Hu, X., Ye, Y., & Yue, X. (2026). *CausalGDP: Causality-Guided Diffusion Policies for Reinforcement Learning*. arXiv preprint arXiv:2602.09207v1.
- Xu, Y., Ganesh, S., Mondal, W. U., Bai, Q., & Aggarwal, V. (2025). *Global Convergence for Average Reward Constrained MDPs with Primal-Dual Actor Critic Algorithm*. 39th Conference on Neural Information Processing Systems (NeurIPS 2025). arXiv preprint arXiv:2505.15138v2.
- Yu, H. (2021). *On Linear Programming for Constrained and Unconstrained Average-Cost Markov Decision Processes with Countable Action Spaces and Strictly Unbounded Costs*. arXiv preprint arXiv:1905.12095v3.
- Zuo, Q., & He, F. (2025). *Ensuring Safety in an Uncertain Environment: Constrained MDPs via Stochastic Thresholds*. arXiv preprint arXiv:2504.04973v3.