Balancing Policy Objectives in Social Security Reform

Arvind Sharma, Aleksandar Tomic, and Lawrence Fulton

Department of Applied Analytics and Economics

Boston College

Author Note

Correspondence concerning this article should be addressed to Arvind Sharma, Department of Applied Analytics and Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467. Email: arvind.sharma@bc.edu

Abstract

Social Security faces structural insolvency by 2033, requiring comprehensive reform balancing competing policy objectives. This study develops an optimization framework integrating Hierarchical Weighted Multi-voting with Mixed Integer Goal Programming to evaluate 142 reform proposals systematically. We employ four Large Language Models to assess each course of action across six policy dimensions: individual cost burden, trust, equity, sustainability, administrative feasibility, and political viability. Monte Carlo simulation with Dirichlet-sampled weights captures evaluative uncertainty across 100 optimization runs, with all solutions satisfying the statutory 3.5% actuarial balance requirement. Results identify progressive revenue enhancement as the dominant reform strategy across all 100 optimization scenarios, with tax base expansion consistently outperforming benefit reduction and privatization alternatives. These findings remain robust across diverse value weightings and policy priorities. The framework demonstrates how structured decision-support tools incorporating artificial intelligence can navigate multidimensional trade-offs in complex policy environments.

Keywords: Social Security Reform, Mixed Integer Goal Programming, Monte Carlo Simulation, Large Language Models, Policy Analysis

Balancing Policy Objectives in Social Security Reform

Issue: Social Security's Structural Crisis

Social Security remains the cornerstone of retirement security in the United States, supporting approximately 90% of Americans aged 65 and older (Social Security Administration, 2024b). In 2023, the program distributed $1.237 trillion in benefits, with 77.8% allocated to retirees and dependents, 11% to disabled workers, and 11.2% to survivors (Social Security Administration, 2024b). However, the Old-Age and Survivors Insurance Trust Fund faces imminent depletion, projected to exhaust reserves by 2033 under intermediate assumptions (Board of Trustees, 2024).

The program's fiscal crisis stems from fundamental structural challenges. In 2023, outlays exceeded income by $70 billion, with costs projected to exceed revenues annually through at least 2098 (Board of Trustees, 2024). The trust fund reserve ratio declined from 188% in 2023 and will reach 84% by 2030 before complete exhaustion (Social Security Administration, 2024a). By statute, the Social Security Administration cannot disburse benefits exceeding available reserves (Lanza & Nicola, 2014), meaning depletion would trigger immediate 21% benefit reductions for all recipients.

Three interconnected factors drive this crisis. First, increasing longevity extends benefit payment periods without proportional contribution increases (Prettner & Canning, 2014). Second, declining labor force participation reduces the worker-to-beneficiary ratio (Fry, 2020). Third, accelerated population aging as baby boomers retire creates unprecedented demographic strain (Center for Retirement Research, 2024). The worker-to-beneficiary ratio has declined from 16.5:1 in 1950 to approximately 2.8:1 today, fundamentally challenging the program's pay-as-you-go financing structure (Attanasio et al., 2007; Börsch-Supan & Schnabel, 1998).

Historical reform efforts illustrate both the feasibility and limitations of comprehensive approaches. The Social Security Amendments of 1983 introduced benefit taxation, gradual retirement age increases, and delayed cost-of-living adjustments (United

States Congress, 1983). The Omnibus Budget Reconciliation Act of 1993 expanded benefit taxation for higher earners (Budget Counsel, 2017). While these reforms extended solvency temporarily, they proved insufficient to address long-term structural imbalances. Simulation studies emphasize the need for deeper modeling that reveals important interactions challenging conventional wisdom (Auerbach & Kotlikoff, 1987; Nishiyama & Smetters, 2007).

Traditional policy analysis struggles to navigate competing objectives inherent in Social Security reform. Any proposal generates effects across multiple dimensions: fiscal sustainability, distributional equity, administrative complexity, political feasibility, and public trust (Diamond & Orszag, 2005). These objectives often conflict—measures enhancing sustainability may impose disproportionate burdens on vulnerable populations, while politically viable reforms may prove insufficient to achieve actuarial balance (Feldstein & Liebman, 2002; Fuster, 2008). Furthermore, reform proposals exhibit complex interdependencies where effectiveness depends on policy combinations rather than individual interventions, as demonstrated in overlapping generations models (Fehr et al., 2008; İmrohoroğlu et al., 2003).

The literature on Social Security reform modeling has evolved substantially, yet important limitations remain. Early studies focused on parametric reforms adjusting basic parameters like tax rates and benefit formulas using deterministic projections (Auerbach & Kotlikoff, 1987). Subsequent research incorporated behavioral responses and general equilibrium effects (Kitao, 2014; Nishiyama & Smetters, 2007), examined privatization proposals through lifecycle frameworks (Fehr, 2008; İmrohoroğlu et al., 2003), and analyzed distributional consequences across income and demographic groups (Fuster, 2008; Gustman & Steinmeier, 2001). More recent work employs sophisticated microsimulation models capturing heterogeneity and uncertainty (Urban Institute, 2020), explores international reform experiences (Börsch-Supan & Schnabel, 1998), and examines policy interactions with health insurance and retirement behavior (İmrohoroğlu et al., 1998; Scholz et al., 2006).

However, as Kitao (2014) and Nishiyama and Smetters (2007) emphasize, analytical models—however sophisticated—cannot fully capture the political, institutional, and behavioral complexities inherent in major policy reforms. These tools provide decision support rather than definitive answers, illuminating trade-offs and identifying robust strategies while acknowledging irreducible uncertainty (Lee & Carter, 1992). Most existing approaches either optimize single objectives or rely on ad hoc weighting schemes that obscure normative assumptions. Our framework advances this literature by explicitly modeling multiple objectives with stochastic weighting, capturing evaluative diversity through AI-driven assessments, and ensuring solutions satisfy binding fiscal constraints. Critically, we position our analysis as complementing rather than replacing democratic deliberation, stakeholder consultation, and expert judgment in the policy formation process.

This complexity demands analytical tools capable of simultaneously considering multiple objectives, incorporating diverse evaluative perspectives, respecting binding constraints, and identifying robust solutions under uncertainty. Our framework addresses these requirements through structured optimization enhanced by artificial intelligence-driven evaluation, offering transparent and reproducible decision support—not prescriptive solutions—for evidence-based policy deliberation.

This study makes three contributions to Social Security reform analysis. First, we develop a novel multi-criteria evaluation framework using Large Language Models as systematic expert panels, providing reproducible assessment with demonstrated inter-rater reliability (weighted kappa 0.601–0.871 for technical constructs) comparable to human expert consensus. This approach advances beyond ad hoc weighting schemes that obscure normative assumptions in existing reform analyses. Second, we demonstrate that progressive revenue enhancement dominates benefit reduction across all tested scenarios, a finding robust to substantial variation in value priorities and policy weights. Third, we provide implementable policy recommendations grounded in constrained optimization that directly address the 2033 trust fund depletion, translating analytical findings into actionable reform strategy.

Model: Optimization Framework and Methods

We develop a comprehensive decision-support framework combining Mixed Integer Goal Programming (Charnes et al., 1961) with Hierarchical Weighted Multi-voting informed by Large Language Model evaluations. This approach builds on established methods in multi-criteria decision analysis (Delbecq & Van de Ven, 1971; Van de Ven & Delbecq, 1974) while incorporating modern AI capabilities (Safaei & Longo, 2024).

Data and Evaluation Structure

Our analysis encompasses 142 distinct reform proposals (Courses of Action, COAs). The first 140 derive from the Social Security Administration's comprehensive evaluation, with actuarial impacts based on the 2024 Trustees Report (Social Security Administration, 2024c). These officially evaluated proposals represent decades of policy development and actuarial analysis, providing a robust foundation for optimization. Two additional privatization proposals (COAs 141–142) represent frequently advocated approaches not formally evaluated by SSA, allowing comparison between traditional financing and fundamental restructuring options. All data and code used in this analysis are openly available at https://github.com/dustoff06/SocialSecurity/.

The 142 COAs span nine policy categories. *Cost of Living Adjustments* (9 policies) modify the inflation indexing mechanism, including proposals to implement chained CPI or reduce adjustment percentages (Social Security Administration, 2008). *Benefits* (51 policies) encompass formula modifications, minimum benefit enhancements, and targeted reductions, addressing both adequacy and fiscal concerns. *Age-Based* changes (14 policies) adjust full retirement ages and early eligibility thresholds, reflecting increased longevity (Prettner & Canning, 2014). *Family Benefits* (9 policies) modify spouse and survivor provisions. *Taxation* (35 policies) includes payroll tax rate changes, taxable maximum adjustments, and coverage expansions—representing the largest category due to revenue generation potential. *Coverage* expansions (10 policies) extend mandatory participation to currently exempt state and local employees. *Equity Investment* (7 policies) allow trust fund diversification beyond Treasury securities. *Benefit Taxation* (5 policies) modify income taxation of benefits. *Privatization* (2 policies) introduce individual retirement accounts partially or fully replacing traditional benefits.

Each COA was evaluated across six policy constructs using four state-of-the-art Large Language Models: ChatGPT-4o (OpenAI, 2024), Claude 3.7 Sonnet (Anthropic, 2024), Gemini 2.5 Pro Experimental (Google DeepMind, 2024), and GitHub Copilot (Microsoft, 2024). These models represent diverse training data, architectures, and evaluative approaches, analogous to expert panels in structured decision-making (Van de Ven & Delbecq, 1974). The six constructs are:

Individual Cost Burden (ICB): Direct financial impact on individuals through tax increases or benefit reductions, emphasizing protection of vulnerable populations. Evaluation considers both magnitude and distribution of burdens, with particular attention to impacts on low-income workers, disabled beneficiaries, and those with limited economic resources.

Trust (T): Alignment with Social Security's foundational principles of earned benefits, universality, defined benefits, and intergenerational solidarity (Zhi et al., 2022). Trust evaluation assesses whether reforms maintain the program's insurance character versus transforming it toward means-tested assistance or private accounts.

Equity (E): Fairness in distributional impacts across income levels, demographic groups, and generations. Equity assessment considers differential effects by race, gender, disability status, and occupation, recognizing that seemingly neutral policies often generate disparate impacts (Social Security Administration, 2023).

Sustainability (S): Long-term fiscal viability and ability to meet obligations indefinitely. Sustainability evaluation encompasses both actuarial balance improvements and resilience to economic and demographic uncertainty.

Administrative Feasibility (AF): Implementation complexity, information requirements, verification systems, and organizational capacity needs. Feasibility assessment recognizes that theoretically superior policies may prove impractical when implementation demands exceed administrative capacity (Demirkiran et al., 2015).

Political Viability (PV): Likelihood of legislative enactment and public acceptance based on historical precedents, stakeholder positions, and public opinion. While political constraints do not determine policy merit, acknowledging them proves essential for practical reform strategy.

Large Language Model Evaluation Process

The LLM evaluation process followed structured protocols inspired by established expert elicitation methods (Delbecq & Van de Ven, 1971). First, a calibration phase aligned model understanding through training on 20 representative COAs selected to span the full range of policy categories and actuarial impacts. Each model received detailed construct descriptions, scoring criteria, and examples of high- and low-scoring proposals. Training materials emphasized systematic, consistent application of criteria while acknowledging legitimate interpretive differences.

Following initial training, preliminary evaluations underwent human expert review to identify systematic biases, misunderstandings, or inconsistencies. Feedback clarified ambiguities and reinforced evaluation standards. This iterative calibration process—common in Delphi techniques (Van de Ven & Delbecq, 1974)—enhanced both scoring quality and cross-model comparability.

Second, comprehensive evaluation generated scores for all 142 COAs across six constructs using -5 to +5 scales, where negative scores indicate adverse effects and positive scores indicate beneficial impacts. Models provided both numerical scores and brief textual rationales explaining assessments, enhancing transparency and facilitating validation.

Third, a reflective feedback mechanism allowed models to revise assessments after reviewing other models' evaluations and rationales. Each model received complete visibility into others' scores for every COA-construct combination. This transparency enabled consideration of alternative perspectives, identification of potential oversights, and revision where justified. Some models demonstrated substantial flexibility in updating assessments based on compelling alternative arguments, while others (particularly Gemini) exhibited relative intransigence, maintaining initial evaluations despite exposure to divergent views.

This variation in revision propensity reflects real-world expert behavior where some analysts readily incorporate new information while others maintain strong priors. Rather than forcing convergence, we preserved evaluative diversity as a methodological asset. In complex policy domains characterized by genuine normative disagreement and analytical uncertainty, variation across informed evaluators provides valuable information about the range of defensible positions (Safaei & Longo, 2024).

The LLM evaluation approach offers distinct advantages over traditional expert elicitation while acknowledging complementary roles. First, it ensures perfect reproducibility—any researcher can regenerate identical evaluations from the same prompts and model versions, addressing replication challenges in policy analysis (Safaei & Longo, 2024). Second, it systematically captures diverse evaluative perspectives through architecturally distinct models trained on different corpora, rather than relying on convenience samples of available experts. Third, the structured feedback mechanism generates documented rationales for each assessment, enabling transparent auditing of evaluative logic. While LLM assessments cannot replace stakeholder input and political judgment in democratic policy formation, they provide consistent, scalable decision support for systematic comparison of reform alternatives under explicit criteria.

Descriptive Statistics and Inter-Rater Reliability

Table 1 presents summary statistics for evaluation scores across LLM assessments and the actuarial effect measure. For each variable, we report minimum, median, mean (standard deviation), maximum, and interquartile range.

<Insert Table 1 about here.>

Evaluation scores range broadly across constructs, with minimum values frequently reaching -5, suggesting strong negative assessments for some COAs. Medians for many criteria cluster near -1 or 0, indicating moderate or slightly negative evaluations across most models. Mean and standard deviation values reveal variability both across and within constructs. Within ICB evaluations, means ranged from -0.90 (ChatGPT) to -0.40 (Copilot), with wide standard deviations around 2, signaling notable differences in perceived burden depending on the model.

Evaluations of Trust are also skewed slightly negative on average, though variability remained substantial. Assessments of Equity were more mixed: some evaluations (Claude, Gemini) reflected positive means, while others (ChatGPT, Copilot) remained negative, suggesting fundamental differences in how models weigh competing equity concerns.

Both Sustainability and Administrative Feasibility constructs showed generally positive evaluations, with Sustainability (Copilot) at 2.42 and Administrative Feasibility (Claude) at 2.63, suggesting many COAs are seen as relatively sustainable and administratively feasible. By contrast, Political Viability evaluations were predominantly negative across all models, indicating that many proposed reforms might struggle to achieve political support even if technically sound. This pattern helps explain why Social Security reform remains "the third rail" of American politics—the tension between technical viability and political feasibility creates persistent gridlock.

The actuarial effect variable had a mean of 0.64% with standard deviation 0.96%, ranging from -1.48% to 4.13%. The interquartile range of 1.07% indicates that while outliers exist, most reforms cluster closely around modestly positive improvements to solvency. Combined with positive means on Sustainability and Administrative Feasibility, this suggests the technical dimension of reform is more tractable than the political dimension.

To assess LLM evaluation consistency, we computed both Spearman rank correlation coefficients and weighted Cohen's kappa statistics for all pairwise comparisons within each construct. Spearman's rho measures monotonic association, capturing whether models agree on relative COA rankings even if absolute score scales differ. Weighted kappa quantifies absolute agreement accounting for the ordinal nature of ratings, assigning partial credit for near-miss agreements rather than treating all disagreements equally.

Results demonstrated substantial inter-model consistency for technically grounded constructs. Individual Cost Burden assessments showed weighted kappa values ranging from 0.601 to 0.828 across model pairs, indicating substantial to almost perfect agreement. Similarly, Trust evaluations exhibited strong coherence (weighted kappa 0.670–0.871), as did Sustainability assessments (0.782–0.861). These high reliability values reflect the relatively objective, quantifiable foundations of these constructs, where clear analytical frameworks guide evaluation.

In contrast, Equity assessments displayed more modest agreement (weighted kappa 0.334–0.670), with several pairwise comparisons showing only fair concordance. This lower consistency likely reflects the inherently contested nature of equity judgments, where different ethical frameworks (utilitarian versus Rawlsian, for example) and empirical assumptions yield divergent conclusions. Administrative Feasibility showed similarly moderate agreement (weighted kappa 0.087–0.363), potentially reflecting varying assumptions about implementation constraints and organizational capacity. Political Viability demonstrated intermediate consistency (weighted kappa 0.546–0.821), suggesting reasonable consensus regarding political constraints despite the judgment-intensive nature of such assessments.

These reliability patterns informed our methodological choices. Rather than down-weighting low-agreement constructs, we preserve diversity through hierarchical probabilistic weighting. The framework treats variation as informative rather than problematic, recognizing that in contested policy domains, disagreement among informed evaluators reflects genuine uncertainty rather than measurement error.

Hierarchical Probabilistic Weighting

To capture uncertainty in policy prioritization and evaluator credibility, we employ two-tier Dirichlet distribution sampling (Gelman et al., 2013). This hierarchical structure generates diverse weight configurations reflecting the range of plausible value systems and interpretive frames relevant to Social Security reform.

At the construct level, weights w_c for constructs $c \in \{ICB, T, E, S, AF, PV\}$ are drawn from a symmetric Dirichlet distribution:

$$\mathbf{w} = (w_{ICB}, w_T, w_E, w_S, w_{AF}, w_{PV}) \sim \text{Dirichlet}(\alpha, \alpha, \alpha, \alpha, \alpha, \alpha) \tag{1}$$

where $\alpha = 2$ for all constructs. The Dirichlet distribution ensures $\sum_c w_c = 1$ and $w_c \geq 0$ for all c, providing a valid probability distribution over construct importance while allowing substantial variation across simulation runs. The choice of $\alpha = 2$ reflects moderate prior uncertainty about relative construct importance. Larger values would concentrate distributions near uniform weighting, while smaller values (especially $\alpha < 1$) would favor extreme allocations. The symmetric parameterization treats all constructs as equally likely to receive high weight a priori, though realized weights vary considerably across draws.

Within each construct c, we sample relative weights for the four LLM evaluations. Let $b_{c,m}$ denote the weight assigned to model $m \in \{1, 2, 3, 4\}$ within construct c. These within-construct weights are also drawn from symmetric Dirichlet distributions:

$$\mathbf{b}_c = (b_{c,1}, b_{c,2}, b_{c,3}, b_{c,4}) \sim \text{Dirichlet}(\beta, \beta, \beta, \beta) \tag{2}$$

where $\beta = 2$ for all models. This structure captures uncertainty about which evaluative perspectives should receive greater influence, modeling the realistic scenario where expert credibility and relevance vary across decision contexts.

The hierarchical weights combine multiplicatively to produce goal-specific weights for each LLM-construct combination. Let $R_{i,c,m}$ denote model m's rating of COA i on construct c. The composite score for COA i is:

$$S_i = \sum_c w_c \left(\sum_m b_{c,m} \cdot R_{i,c,m} \right) \tag{3}$$

This formulation first computes a weighted average of model scores within each construct, then aggregates across constructs using construct-level weights. The multiplicative structure preserves interaction effects between construct importance and model influence, avoiding the artificial separation that would result from additive aggregation. Additionally, the actuarial effect of each COA enters the optimization directly as a hard constraint rather than through the composite scoring function, recognizing actuarial balance as a legally mandated threshold requirement rather than a soft objective amenable to trade-offs.

Mixed Integer Goal Programming Formulation

The optimization model identifies COA portfolios maximizing total weighted score subject to binding constraints on actuarial balance, portfolio size, and policy restrictions. The formulation balances multiple objectives while ensuring all solutions satisfy statutory requirements.

Let $x_i \in \{0, 1\}$ indicate whether COA i is selected ($x_i = 1$) or not ($x_i = 0$) for $i \in \{1, 2, \ldots, 142\}$. Additionally, let $d_i^- \geq 0$ represent negative deviation from target performance for COA i, capturing the extent to which selected policies fall short of aspirational goals.

The objective maximizes total weighted contribution minus penalized shortfalls:

$$\text{maximize} \quad Z = \sum_{i=1}^{142} S_i \cdot x_i - \lambda \sum_{i=1}^{142} d_i^- \tag{4}$$

where λ is a penalty weight drawn from Uniform(0.2, 0.7), reflecting variability in tolerance for goal underachievement. Lower values emphasize maximizing total weighted score with less concern for specific target achievement, while higher values prioritize meeting individual COA performance targets even if reducing overall portfolio score.

The complete formulation includes several critical constraints. The actuarial balance requirement ensures the selected portfolio achieves at least 3.5% of payroll in long-term balance:

$$\sum_{i=1}^{142} A_i \cdot x_i \geq 3.5 \tag{5}$$

where A_i denotes the actuarial effect of COA i. This constraint is binding in all runs, ensuring legal compliance (Lanza & Nicola, 2014).

The privatization restriction limits selection to at most one privatization proposal:

$$\sum_{i=141}^{142} x_i \leq 1 \tag{6}$$

reflecting that privatization represents fundamental program redesign incompatible with traditional financing.

Portfolio cardinality is bounded by:

$$\sum_{i=1}^{142} x_i \leq N_{\max} \tag{7}$$

where $N_{\max}$ is drawn uniformly from $\{1, 2, 3, 4, 5, 6\}$, reflecting political and administrative realism that comprehensive reforms combine multiple policies but excessively complex packages become difficult to communicate and implement.

Goal deviation tracking captures performance shortfalls:

$$S_i \cdot x_i - d_i^- \leq T_i \quad \forall i \tag{8}$$

where T_i represents aspirational targets. These soft constraints allow selecting COAs falling short of idealized performance, with penalties adjusted by λ. Complete mathematical specifications and derivations appear in Online Appendix A.

Simulation Design and Implementation

We executed 100 Monte Carlo iterations, each sampling fresh construct weights, LLM weights, penalty parameters, and portfolio size limits. For each simulation run $r = 1, \ldots, 100$: (1) Sample construct weights from Dirichlet(2,2,2,2,2,2); (2) For each construct, sample model weights from Dirichlet(2,2,2,2); (3) Compute composite scores via Equation (3); (4) Sample penalty parameter from Uniform(0.2, 0.7); (5) Sample maximum portfolio size from Discrete Uniform$\{1, 2, 3, 4, 5, 6\}$; (6) Solve optimization model using lpSolveAPI (Berkelaar et al., 2024; R Core Team, 2024); (7) Record selected COAs, objective value, actuarial balance, and diagnostic statistics.

To evaluate whether 100 runs provide sufficient sampling, we computed congruence scores based on Jaccard similarity across all runs for each $N_{\max}$. Results indicated perfect selection stability for $N_{\max} \in \{1, 2, 3, 4\}$ and high stability for $N_{\max} \in \{5, 6\}$, with average congruence scores of 0.85. These findings suggest the solution space is well-behaved and sufficiently explored at the current sampling level, supporting the use of 100 runs for stable policy selection analysis. Outputs were aggregated to assess selection frequencies, solution stability, construct satisfaction patterns, and co-selection correlations. Complete simulation algorithm and convergence diagnostics appear in Online Appendix B.

Validation: Framework Reliability Assessment

We validate the framework through solution stability analysis, constraint verification, sensitivity analysis, and comparison with alternative methodologies.

Solution Convergence and Stability

Selection frequency distributions demonstrate robust convergence. The top 15 COAs appeared in 38%–86% of runs, indicating consistent performance across diverse weighting scenarios. Conversely, 68 COAs were never selected in any run, indicating they are dominated by superior alternatives across all evaluated preference structures. This clear differentiation between consistently selected and never-selected policies validates that the optimization effectively discriminates based on multi-criteria performance.

Jaccard similarity analysis confirms solution stability. For portfolio sizes $N_{\max} \in \{1, 2, 3, 4\}$, solutions exhibited perfect consistency—when the same maximum was drawn, identical COA sets were selected regardless of weight variation. This perfect stability for smaller portfolios indicates strong dominance relationships among top-performing policies. For larger portfolios ($N_{\max} \in \{5, 6\}$), average Jaccard similarity exceeded 0.85, indicating substantial overlap despite greater combinatorial possibilities.

The high stability across diverse weight configurations suggests our framework identifies genuinely robust solutions rather than artifacts of particular assumptions.

Constraint Satisfaction and Feasibility

All 100 optimization runs successfully identified feasible solutions satisfying binding constraints, demonstrating computational reliability. The actuarial balance constraint achieved 100% compliance, with no run producing portfolios below the 3.5% threshold. Actual achieved balances averaged 4.2% (standard deviation 0.6%), safely exceeding the minimum while avoiding excessive taxation that would impose unnecessary economic burdens.

Portfolio size constraints were always binding—every run selected exactly $N_{\max}$ COAs, indicating the optimizer fully utilized available selection capacity. This pattern suggests even small portfolios can achieve actuarial balance while satisfying multiple objectives, but larger portfolios perform better on composite weighted scores by incorporating diverse policy mechanisms. The privatization restriction never prevented optimal solution identification. In the 8 runs where privatization COAs were selected, only one appeared per portfolio as required. The vast majority of runs (92%) excluded both privatization proposals, finding superior alternatives among traditional financing reforms.

Sensitivity Analysis and Robustness

We examined how solution characteristics vary with key parameters. Penalty weight λ showed modest influence—higher penalties slightly increased selection of COAs with strong across-the-board performance, while lower penalties favored specialists excelling on specific constructs. However, the top-tier COAs remained dominant across the full λ range, indicating their selection is robust to penalty specification.

Construct weights exhibited stronger effects, as expected. When Trust and Equity received above-median weights, progressive benefit formula modifications gained prominence. When Sustainability and Political Viability dominated, tax increases and age adjustments appeared more frequently. However, certain COAs (particularly tax base expansions) maintained high selection frequencies across nearly all weight configurations, indicating genuine robustness to value system variation.

Maximum portfolio size naturally affected diversity. Larger $N_{\max}$ values enabled inclusion of complementary policies addressing multiple dimensions, while small portfolios concentrated on high-impact core reforms. Interestingly, objective values did not increase monotonically with $N_{\max}$, suggesting diminishing returns to portfolio expansion beyond 4–5 policies. This finding supports reform strategies focusing on a manageable set of high-impact interventions rather than attempting comprehensive omnibus packages.

Comparison with Alternative Methodologies

We compared results against three alternative approaches to validate our framework's value-added. First, equal-weighted aggregation (treating all constructs and models identically) produced similar top-tier COAs but narrower solution sets, suggesting our stochastic approach identifies a broader range of defensible alternatives. Second, single-objective optimization (maximizing actuarial impact only) selected portfolios dominated by large tax increases, ignoring equity and feasibility concerns. These solutions achieved higher actuarial balance but performed poorly on other dimensions, illustrating the importance of multi-criteria optimization (Fehr, 2008).

Third, benefit-reduction-focused portfolios (constraining revenue measures while emphasizing benefit cuts) performed poorly on Trust and Individual Cost Burden dimensions while achieving inferior actuarial results. This finding aligns with research emphasizing progressive taxation's advantages over regressive benefit cuts (Diamond & Orszag, 2005). These comparisons validate that our multi-criteria, probabilistically weighted framework generates solutions better balancing diverse objectives than simpler alternatives.

Findings: Optimal Reform Strategies

Tax-based revenue enhancement emerges as the dominant reform strategy, appearing in 73 of 100 optimization runs (73%), with complete taxable maximum elimination (COA 90) selected in 61% of scenarios. Progressive benefit enhancements protecting vulnerable populations appear in 86% of runs despite negative actuarial impact, indicating their necessity for maintaining program legitimacy across diverse value systems. By contrast, privatization appears in only 8% of runs and across-the-board benefit cuts in fewer than 15%. This dominance pattern persists across all tested weight configurations, indicating robustness to value system variation rather than sensitivity to particular normative assumptions.

Most Frequently Selected Policies

Table 2 presents the 15 most frequently selected COAs, their policy categories, selection frequencies, and actuarial impacts. Tax-related measures generating substantial positive actuarial effects dominate the top ranks, appearing with significantly higher frequency than benefit reduction alternatives. Benefit modifications appearing in optimal portfolios split between progressive reforms protecting vulnerable populations (COA 32: 86% selection frequency, -0.13% actuarial impact) and targeted adjustments to high-earner benefits (COA 17: 45% selection frequency, +0.67% actuarial impact). This pattern demonstrates that fiscal sustainability and distributional equity function as complements rather than competing objectives when reform portfolios combine progressive revenue enhancement with targeted vulnerability protections.

<Insert Table 2 about here.>

Dominant Policy Categories and Mechanisms

Tax measures constitute 47% of all COA selections across 100 runs, far exceeding any other category. This dominance reflects both the magnitude of revenue generation potential and favorable performance across multiple constructs when designed progressively. Three distinct tax approaches appear prominently:

Eliminating or raising the taxable maximum (COAs 90, 96, 103): Removing or increasing the earnings cap appears in 61% of runs for complete elimination (COA 90), generating 3.95% actuarial improvement while enhancing progressivity. This class of reforms performs exceptionally well on Equity and Sustainability while maintaining reasonable Individual Cost Burden scores when concentrated on highest earners. Alternative approaches creating "donut holes" (taxing earnings above thresholds while exempting middle ranges) show similar appeal, combining revenue generation with perceived fairness.

Increasing payroll tax rates (COAs 85, 87): Direct rate increases generate large actuarial improvements—COA 85's phased increase to 19.4% yields 4.13% and appears in 73% of runs despite imposing broadly distributed burdens. These proposals perform well when weighted toward Sustainability, though their selection typically occurs alongside progressive benefit enhancements to maintain acceptable Individual Cost Burden and Equity scores. The high selection frequency indicates acceptability when combined with measures protecting vulnerable populations.

Coverage expansion (COAs 119, 121): Extending mandatory Social Security coverage to state and local government employees improves actuarial balance while enhancing equity through broader risk pooling (Social Security Administration, 2024c). COA 121 (covering only new hires rather than current employees) achieves 52% selection frequency, balancing sustainability gains against political feasibility concerns about disrupting existing pension arrangements.

Benefits modifications represent 26% of selections, with progressive adjustments substantially dominating regressive cuts. The special minimum benefit enhancement (COA 32) appears in 86% of runs despite negative actuarial impact (-0.13%), reflecting exceptionally strong performance on Individual Cost Burden, Equity, and Trust dimensions. This policy provides essential protection for low-lifetime-earnings workers, maintaining program legitimacy. Across-the-board benefit increases (COA 31) appear in 63% of runs when combined with substantial revenue measures, demonstrating that enhanced adequacy remains compatible with fiscal sustainability in well-designed portfolios.

Progressive benefit formula modifications (COAs 17, 20) that reduce replacement rates for highest earners while protecting low and middle earners appear in 21%–45% of runs. These reforms simultaneously improve actuarial balance and enhance equity, though their relatively lower selection frequency compared to revenue measures suggests preference for explicit progressive taxation over implicit benefit redistribution. This pattern aligns with public opinion research showing greater acceptance of tax-side progressivity (Diamond & Orszag, 2005).

Age-related proposals show mixed performance. Gradually increasing full retirement age to 69 (COA 67) appears in 38% of runs, providing 0.81% actuarial improvement. However, more aggressive age increases face resistance due to equity concerns regarding workers in physically demanding occupations and demographic groups with below-average life expectancy. The moderate selection frequency suggests age increases serve as useful supplementary measures but rarely constitute primary reform strategies in optimal portfolios. This finding reflects important distributional considerations emphasized in the literature on health, mortality, and retirement (Cutler et al., 2011).

Investment diversification proposals (COAs 129–133) demonstrate surprising appeal despite implementation complexity. COA 129, allocating 40% of trust fund assets to equities, appears in 31% of runs with 0.95% actuarial impact based on expected return differentials. These proposals score well on Sustainability due to higher projected returns, though lower Administrative Feasibility and Political Viability ratings limit their selection frequency. The significant minority inclusion suggests genuine policy interest merits further exploration, particularly regarding governance structures and risk management protocols.

Notably Absent Policies

Privatization proposals (COAs 141–142) appear in only 8% of runs, constrained by poor performance on Trust, Equity, and Administrative Feasibility dimensions despite potentially positive sustainability effects under optimistic return assumptions. This finding aligns with research emphasizing privatization's implementation challenges and distributional concerns (Fehr et al., 2008). The low selection frequency validates that within our multi-criteria framework, privatization underperforms traditional financing reforms.

Across-the-board benefit cuts appear in fewer than 15% of runs, underperforming on nearly all dimensions except Sustainability. Their poor showing on Trust, Equity, and Individual Cost Burden dimensions renders them dominated by alternative approaches achieving comparable actuarial improvements through progressive revenue enhancement. Highly complex administrative reforms requiring extensive new infrastructure rarely appear despite favorable actuarial impacts, underscoring Administrative Feasibility as a genuine binding constraint rather than merely theoretical consideration.

Portfolio Characteristics and Synergies

Selected portfolios typically combine 3–4 complementary policies spanning multiple categories rather than relying on single interventions. This finding validates comprehensive reform approaches advocated in the policy literature (Diamond & Orszag, 2005; Urban Institute, 2020). Actuarial balance averaged 4.2% (standard deviation 0.6%), safely exceeding the 3.5% statutory minimum while avoiding excessive taxation.

Correlation analysis reveals important complementarities and substitution relationships. Enhanced minimum benefits (COA 32) and taxable maximum elimination (COA 90) show positive co-selection correlation (0.42), suggesting these reforms form natural progressive complements balancing equity across revenue and benefit dimensions. Similarly, coverage expansion (COA 121) and progressive benefit formulas (COA 17) exhibit positive correlation (0.38), both enhancing program universality and progressivity.

Conversely, large payroll tax increases (COA 85) and taxable maximum elimination (COA 90) show negative correlation (-0.31), reflecting substitution—portfolios typically include one major revenue measure rather than combining multiple large-scale tax increases that would impose excessive burdens. Age increases (COA 67) and across-the-board benefit increases (COA 31) similarly demonstrate negative correlation (-0.38), embodying opposing philosophical approaches to generational equity and benefit adequacy. These patterns inform reform strategy by identifying natural policy bundles and highlighting incompatible combinations.

Advisory: Evidence-Based Reform Strategy

Our optimization framework identifies specific policy combinations that consistently achieve fiscal sustainability while optimizing equity, trust, and political feasibility. Based on selection frequencies across 100 scenarios with diverse value weightings, we recommend a four-policy portfolio that substantially exceeds the 3.5% actuarial requirement while demonstrating superior performance on distributional and legitimacy dimensions. The following sections detail core revenue strategy, progressive benefit adjustments, and implementation considerations.

Core Revenue Strategy

Primary Recommendation: Eliminate or substantially raise the taxable maximum. COA 90 (complete elimination) generates 3.95% actuarial improvement while appearing in 61% of optimal solutions across all weighting scenarios. This measure enhances progressivity by requiring higher earners to contribute proportionally more, directly addressing the regressive structure of current payroll taxation where earnings above $168,600 (2024) escape taxation. Implementation should be phased over 10 years to moderate labor market disruption and allow high earners to adjust financial planning (Diamond & Orszag, 2005).

Alternative Approaches: If complete elimination faces insurmountable political resistance, implement COA 96 (uncap maximum while exempting $400k–$500k range), generating 2.91% improvement with potentially greater political viability by creating a "donut hole" that concentrates new taxation on very highest earners. Alternatively, implement COA 103 (tax earnings above $250k individual/$500k joint) for 1.55% improvement with more targeted impact and clearer connection to ability-to-pay principles.

Supplementary Revenue: If taxable maximum modification proves insufficient to reach 3.5% threshold or if greater fiscal buffer is desired, consider moderate payroll tax rate increase (COA 87: 3.8 percentage point increase yielding 3.15%). This broadly distributed approach showed 58% selection frequency and performs well when combined with progressive benefit enhancements that offset burdens on lower earners. Phase implementation gradually to moderate macroeconomic impacts and distribute adjustment costs across cohorts (Attanasio et al., 2007).

Progressive Benefit Adjustments

Protect Vulnerable Populations: Implement COA 32 (reconfigured special minimum benefit) providing minimum replacement rate of 125% of federal poverty level for workers with 30+ years of coverage. This policy appeared in 86% of solutions despite -0.13% actuarial cost, reflecting exceptional performance on Individual Cost Burden, Equity, and Trust dimensions. The proposal provides essential protection for low-lifetime-earnings workers, maintaining program legitimacy and demonstrating commitment to adequacy alongside sustainability (Social Security Administration, 2008).

Enhance Adequacy: Consider COA 31 (5% across-the-board benefit increase) when revenue measures generate sufficient actuarial surplus above 3.5% threshold. This proposal appeared in 63% of runs, demonstrating compatibility with fiscal sustainability when paired with substantial revenue enhancements. Across-the-board increases maintain benefit structure simplicity while providing meaningful adequacy improvements, particularly important given stagnant replacement rates for median earners over recent decades.

Progressive Formula Adjustment: Implement COA 17 (progressive benefit formula modification) reducing replacement rates for highest earners while protecting low and middle earners. This generates 0.67% actuarial improvement while enhancing equity. Specifically, adjust bend points to provide lower marginal replacement for earnings above approximately $100,000 in current dollars, indexed to wage growth. Such modifications maintain earned-benefit structure while enhancing progressivity more transparently than indirect approaches.

Coverage and Administrative Reforms

Expand Coverage: Extend mandatory Social Security coverage to newly hired state and local government employees (COA 121), appearing in 52% of optimal portfolios. This generates 0.29% actuarial improvement while improving risk pooling and equity by incorporating a historically privileged group into universal social insurance (Social Security Administration, 2024c). Implementation should include transition provisions for affected employers and employees, allowing continued participation in existing pension plans for current workers while requiring new hires to participate in Social Security. Technical assistance to state and local governments will facilitate smooth transitions.

Consider Modest Age Adjustment: If additional actuarial improvement is needed beyond revenue and benefit adjustments, gradually increase full retirement age to 68 by 2040, indexed to longevity thereafter (modified COA 67). Critically, pair age increases with enhanced protections for workers in physically demanding occupations through liberalized disability qualifications, expanded unemployment insurance for older workers, and targeted benefit enhancements for those claiming early retirement due to health or job loss (Cutler et al., 2011). Without such protections, age increases impose severe hardship on vulnerable populations and undermine equity objectives.

Implementation Strategy and Sequencing

Reform implementation requires careful sequencing and stakeholder engagement to build political support while managing economic transitions. Based on optimization results demonstrating consistent superiority across 100 scenarios, we recommend a phased approach incorporating four key principles:

Phased Implementation: Introduce changes gradually over 10–15 years, allowing workers to adjust retirement planning and minimizing economic disruption. Phase-in periods should vary by policy—immediate implementation for coverage expansion affecting only new hires, but longer transitions for tax changes affecting all workers. Gradual implementation reduces macroeconomic volatility and distributes adjustment costs more equitably across generations (Auerbach & Kotlikoff, 1987).

Stakeholder Engagement: Address concerns of affected groups through targeted protections and transparent communication. Particular attention should focus on age increases (protecting workers in physically demanding occupations), coverage expansion (ensuring state and local government fiscal sustainability), and tax increases (demonstrating how progressive design protects lower and middle earners). Robust public education campaigns emphasizing reform necessity, design principles, and specific individual impacts will build political coalitions supporting comprehensive action (Diamond & Orszag, 2005).

Sunset Provisions: Include periodic review mechanisms enabling course correction if actuarial projections prove inaccurate or economic conditions change dramatically. Automatic stabilizer provisions linking benefit adjustments or tax rates to trust fund ratios or demographic indicators could reduce need for repeated legislative interventions while maintaining long-term sustainability. Such provisions should include sufficient buffers preventing short-term fluctuations from triggering unnecessary adjustments.

Administrative Investment: Allocate resources for technological infrastructure supporting accurate earnings tracking, benefit calculations, and compliance monitoring. Coverage expansion and progressive taxation modifications require enhanced data systems and verification capabilities. SSA modernization investments will ensure implementation feasibility while improving service delivery for all beneficiaries. Adequate administrative funding proves essential for successful reform implementation (Demirkiran et al., 2015).

Recommended Portfolio

The optimization framework identifies the following four-policy portfolio as dominant, selected with high frequency across diverse value weightings and demonstrating superior performance on equity, sustainability, and trust dimensions:

1. Eliminate taxable maximum or raise to cover 90% of aggregate earnings (COA 90 or variant): +3.95% actuarial

2. Enhance special minimum benefit for low-lifetime earners (COA 32): -0.13% actuarial
3. Cover newly hired state and local government employees (COA 121): +0.29% actuarial
4. Progressive benefit formula adjustment targeting highest earners (COA 17): +0.67% actuarial

This portfolio achieves 4.78% actuarial balance, exceeding the 3.5% statutory requirement by 37% and providing substantial buffer against projection uncertainty. The combination of progressive revenue measures (COA 90, 17) with vulnerability protections (COA 32) and coverage expansion (COA 121) optimizes across all six policy dimensions while maintaining political feasibility through targeted rather than universal burden increases.

This portfolio emphasizes progressive taxation over benefit cuts, protects vulnerable populations through enhanced minimums, expands coverage for improved equity and universality, and employs progressive benefit formula adjustments targeting highest earners. Compared to alternatives emphasizing benefit reductions, privatization, or regressive revenue measures, this portfolio demonstrates superior performance on trust, equity, and political viability metrics while achieving comparable or better sustainability outcomes. The approach aligns with successful reform precedents internationally and domestically (Diamond & Orszag, 2005).

Limitations and Future Directions

Our framework analyzes reforms under 2024 Trustees Report assumptions, which represent the authoritative baseline for Social Security planning. While these projections incorporate uncertainty ranges, major deviations— particularly in immigration policy or labor force participation—would alter optimal portfolios. This limitation affects all actuarial analysis; our contribution is systematic comparison of alternatives under consistent assumptions rather than prediction of future outcomes. Economic volatility, unexpected demographic shifts, technological disruptions affecting labor markets, or political realignments could alter optimal portfolios. The 2024 projections assume specific fertility rates, mortality trends, immigration patterns, labor force participation rates, wage growth, and interest rates. Deviations from these assumptions—particularly regarding immigration policy or labor force participation—could substantially affect both baseline projections and reform effectiveness (Board of Trustees, 2024).

The framework should inform rather than replace democratic deliberation and stakeholder consultation. Optimization provides systematic analysis of trade-offs and identifies robust solutions, but ultimate reform decisions require political judgment incorporating values, priorities, and implementation constraints beyond model scope. Future research should extend this framework in several directions.

First, incorporate dynamic programming approaches for sequential reform opportunities, recognizing that policy adjustments occur over time rather than as one-shot interventions. Sequential models could address timing considerations, learning effects, and path dependencies in reform implementation (Attanasio et al., 2007).

Second, integrate behavioral economic modeling for labor supply responses, retirement timing decisions, and savings adjustments. Current analysis uses SSA actuarial estimates assuming specific behavioral responses, but more sophisticated microsimulation incorporating heterogeneous behavioral parameters would enhance precision (Urban Institute, 2020).

Third, extend the framework to other social insurance programs including Medicare, unemployment insurance, and disability programs, exploring potential coordination and integration opportunities. Comprehensive social insurance reform requires systemic perspective recognizing interconnections across programs (Fehr, 2008).

Fourth, develop real-time updating capabilities incorporating new actuarial projections, economic data, and policy evaluations as they become available. Adaptive frameworks maintaining current analysis would enhance policy relevance.

Finally, expand evaluative diversity by incorporating human expert assessments alongside LLM evaluations, comparing and calibrating AI and human judgment. Such comparison would enhance validation while exploring optimal combinations of human and artificial intelligence in policy analysis (Safaei & Longo, 2024).

Concluding Observations

The 2033 depletion deadline demands urgent action. Our optimization framework demonstrates that balanced, comprehensive reform satisfying multiple objectives while respecting fiscal constraints remains achievable. Progressive revenue enhancement, particularly through tax base expansion, emerges as the most robust policy direction across diverse value systems and weighting assumptions, supporting both fiscal sustainability and distributional equity.

The framework's identification of consistent patterns despite substantial parameter variation suggests genuine policy insights rather than artifacts of particular assumptions. The dominance of progressive taxation, strong performance of vulnerability protections, and poor showing of privatization and across-the-board cuts reflect fundamental trade-offs in Social Security reform that persist across evaluative perspectives.

Policymakers should prioritize these evidence-based strategies in reform deliberations, recognizing that delay increases both the magnitude of necessary adjustments and the political difficulty of implementation. The longer reform is postponed, the more abrupt and disruptive required changes become, potentially forcing suboptimal policies that current analysis identifies as dominated alternatives. Early action provides flexibility for gradual implementation, equitable burden distribution, and course correction based on experience.

Progressive revenue enhancement emerges as the dominant Social Security reform strategy under systematic multi-criteria optimization. This finding— robust across 100 scenarios with diverse value weightings—challenges reform proposals emphasizing benefit cuts or privatization. Specifically, complete elimination of the taxable maximum combined with enhanced minimum benefits for vulnerable populations outperforms all alternative combinations, appearing in 61% and 86% of optimal portfolios respectively.

The optimization framework demonstrates that AI-augmented policy analysis can identify robust solutions in contested domains where traditional methods yield ambiguous guidance. Policymakers facing the 2033 depletion deadline should prioritize tax base expansion combined with progressive benefit adjustments, as these strategies consistently satisfy fiscal constraints while optimizing equity and trust objectives. Delay increases both adjustment magnitude and political difficulty; early action enables gradual implementation and course correction based on experience.

References

Anthropic. (2024). Claude 3.7 sonnet [Accessed May 14, 2025].

Attanasio, O., Kitao, S., & Violante, G. L. (2007). Global demographic trends and social security reform. *Journal of Monetary Economics*, *54*(1), 144–198. https://doi.org/10.1016/j.jmoneco.2006.12.002

Auerbach, A. J., & Kotlikoff, L. J. (1987). *Dynamic fiscal policy*. Cambridge University Press.

Berkelaar, M., Eikland, K., & Notebaert, P. (2024). *Lpsolveapi: R interface to lp_solve* [R package version 5.5.2.0-17.11].

Board of Trustees. (2024). *The 2024 annual report of the board of trustees of the federal old-age and survivors insurance and federal disability insurance trust funds* (tech. rep.). Social Security Administration.

Börsch-Supan, A., & Schnabel, R. (1998). Pension reform in germany: The impact on retirement decisions. *FinanzArchiv: Public Finance Analysis*, *54*(3), 393–421.

Budget Counsel. (2017). *The omnibus budget reconciliation act of 1993* (tech. rep.).

Center for Retirement Research. (2024). *Population aging and social security solvency* (tech. rep.). Boston College.

Charnes, A., Cooper, W. W., & Ferguson, R. (1961). Management models and industrial applications of linear programming. *Management Science*, *4*(1), 38–91.

Cutler, D. M., Liebman, J. B., & Smyth, S. (2011). How does your kindergarten classroom affect your earnings? evidence from project star. *Quarterly Journal of Economics*, *126*(4), 1593–1660.

Delbecq, A. L., & Van de Ven, A. H. (1971). A group process model for problem identification and program planning. *Journal of Applied Behavioral Science*, *7*(4), 466–492.

Demirkiran, I., Misra, S., & Wang, Y. (2015). Optimization in medicare reimbursement using integer programming. *Health Care Management Science*, *18*, 326–341.

Diamond, P. A., & Orszag, P. R. (2005). *Saving social security: A balanced approach*. Brookings Institution Press.

Fehr, H. (2008). Cge modeling of social security reforms. *Journal of Policy Modeling*, *30*(3), 345–358. https://doi.org/10.1016/j.jpolmod.2008.01.004

Fehr, H., Habermann, C., & Kindermann, F. (2008). Social security reform with uninsurable income risk and endogenous borrowing constraints. *Journal of Policy Modeling*, *30*(4), 637–651. https://doi.org/10.1016/j.jpolmod.2007.06.002

Feldstein, M., & Liebman, J. B. (2002). Social security. *Handbook of Public Economics*, *4*, 2245–2324.

Fry, R. (2020). *The pace of boomer retirements has accelerated in the past year* (tech. rep.). Pew Research Center.

Fuster, L. (2008). Distributional effects of the transition to a fully funded pension system. *Journal of Policy Modeling*, *30*(4), 629–636. https://doi.org/10.1016/j.jpolmod.2007.06.003

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press.

Google DeepMind. (2024). Gemini 2.5 pro experimental [Accessed May 14, 2025].

Gustman, A. L., & Steinmeier, T. L. (2001). Social security provisions and the labor force participation of older workers. *Journal of Labor Economics*, *19*(3), 531–555.

İmrohoroğlu, A., İmrohoroğlu, S., & Joines, D. H. (1998). The effect of tax-favored retirement accounts on capital accumulation. *American Economic Review*, *88*(4), 749–768.

İmrohoroğlu, A., İmrohoroğlu, S., & Joines, D. H. (2003). Social security reform: A quantitative analysis. *Journal of Economic Dynamics and Control*, *27*(11–12), 2073–2104. https://doi.org/10.1016/S0165-1889(02)00101-7

Kitao, S. (2014). Fiscal sustainability in japan: What to tackle. *Journal of the Japanese and International Economies*, *34*, 33–58. https://doi.org/10.1016/j.jjie.2014.08.001

Lanza, S., & Nicola, D. (2014). Legal restrictions on social security benefit payments. *Social Security Bulletin*, *74*(3).

Lee, R. D., & Carter, L. R. (1992). Modeling and forecasting us mortality. *Journal of the American Statistical Association*, *87*(419), 659–671.

Microsoft. (2024). Github copilot [Accessed May 14, 2025].

Nishiyama, S., & Smetters, K. (2007). Analyzing an aging population—a dynamic general equilibrium approach. *Journal of Policy Modeling*, *29*(4), 607–625. https://doi.org/10.1016/j.jpolmod.2007.05.001

OpenAI. (2024). Chatgpt-4o: Openai's omnimodal ai model [Accessed May 14, 2025].

Prettner, K., & Canning, D. (2014). Increasing life expectancy and optimal retirement in general equilibrium. *Economic Theory*, *56*, 191–217. https://doi.org/10.1007/s00199-013-0765-0

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Safaei, M., & Longo, J. (2024). The end of the policy analyst? testing the capability of artificial intelligence to generate plausible, persuasive, and useful policy analysis. *Digital Government: Research and Practice*, *5*(1), 1–35. https://doi.org/10.1145/3604570

Scholz, J. K., Seshadri, A., & Khitatrakun, S. (2006). Are americans saving optimally for retirement? *Journal of Political Economy*, *114*(4), 607–643.

Social Security Administration. (2008). Distributional effects of reducing the cost-of-living adjustments [Accessed May 14, 2025].

Social Security Administration. (2023). Ssa's equity action plan 2023 update [Accessed May 14, 2025].

Social Security Administration. (2024a). Oasi trust fund, a social security fund [Accessed May 14, 2025].

Social Security Administration. (2024b). Reports, facts, and figures [Accessed May 14, 2025].

Social Security Administration. (2024c). Summary of provisions that would change the social security program [Accessed May 14, 2025].

United States Congress. (1983). Social security amendments of 1983 [Accessed May 14, 2025].

Urban Institute. (2020). Exploring social security reform options [Accessed May 14, 2025].

Van de Ven, A. H., & Delbecq, A. L. (1974). The effectiveness of nominal, delphi, and interacting group decision making processes. *Academy of Management Journal*, *17*(4), 605–621. https://doi.org/10.5465/255641

Zhi, K., Tan, Q., & Chen, S. (2022). How does social security fairness predict trust in government? *International Journal of Environmental Research and Public Health*, *19*(11), 6867. https://doi.org/10.3390/ijerph19116867

Table 1

Summary Statistics for Evaluation Scores

Variable	Min	Median	*M*	*SD*	Max	IQR
ICB1 (ChatGPT)	−5.00	−2.00	−0.90	1.98	4.00	2.75
ICB2 (Claude)	−5.00	−1.00	−0.82	2.29	3.00	4.75
ICB3 (Gemini)	−5.00	−1.00	−0.75	1.37	2.00	2.00
ICB4 (Copilot)	−5.00	0.00	−0.40	2.08	3.00	4.00
T1 (ChatGPT)	−4.00	−1.00	−0.94	1.84	5.00	2.00
T2 (Claude)	−5.00	−1.00	−0.54	2.02	3.00	3.00
T3 (Gemini)	−5.00	0.00	−0.81	1.39	2.00	2.00
T4 (Copilot)	−5.00	−2.00	−0.77	2.45	4.00	4.75
E1 (ChatGPT)	−4.00	0.00	−0.02	2.18	5.00	4.00
E2 (Claude)	−4.00	2.00	0.82	2.24	4.00	4.00
E3 (Gemini)	−4.00	0.00	0.46	1.28	3.00	1.00
E4 (Copilot)	−4.00	−2.00	−0.58	1.97	3.00	3.00
S1 (ChatGPT)	−3.00	2.00	1.22	2.12	4.00	3.00
S2 (Claude)	−5.00	2.00	1.58	2.33	5.00	5.00
S3 (Gemini)	−5.00	2.00	1.23	1.52	4.00	2.00
S4 (Copilot)	−2.00	3.00	2.42	1.81	5.00	2.00
AF1 (ChatGPT)	−3.00	2.00	1.87	1.57	5.00	2.00
AF2 (Claude)	−3.00	3.00	2.63	1.01	5.00	1.00
AF3 (Gemini)	−4.00	0.00	0.49	1.10	2.00	1.00
AF4 (Copilot)	−4.00	2.50	1.60	2.28	5.00	3.00
PV1 (ChatGPT)	−5.00	−1.00	−0.80	2.26	5.00	4.00
PV2 (Claude)	−5.00	−2.00	−1.01	2.42	3.00	5.00
PV3 (Gemini)	−5.00	−1.00	−0.89	1.26	2.00	2.00
PV4 (Copilot)	−5.00	−1.00	−0.34	2.59	4.00	6.00
SS Effect	−1.48	0.37	0.64	0.96	4.13	1.07

Note. ICB = Individual Cost Burden; T = Trust; E = Equity; S = Sustainability; AF = Administrative Feasibility; PV = Political Viability; SS Effect = actuarial impact as percentage of payroll.

Table 2

Most Frequently Selected Courses of Action

COA	Description	Category	% Runs	Actuarial Effect
32	Reconfigure special minimum benefit	Benefits	86%	-0.13%
85	Increase payroll tax to 15.9%, then 19.4%	Tax	73%	4.13%
31	Increase worker benefit by 5%	Benefits	63%	-0.63%
90	Eliminate taxable maximum entirely	Tax	61%	3.95%
87	Increase payroll tax by 3.8 percentage points	Tax	58%	3.15%
121	Cover newly hired state/local employees	Coverage	52%	0.29%
17	Implement progressive benefit formula	Benefits	45%	0.67%
103	Tax earnings above $250k/$500k	Tax	43%	1.55%
67	Increase full retirement age to 69	Age-Based	38%	0.81%
129	Invest 40% of trust fund in equities	Investment	31%	0.95%
96	Uncap maximum, exempt $400k–$500k	Tax	28%	2.91%
136	Tax benefits like pension income	Benefit Tax	24%	1.16%
20	Reduce benefits for top 25% earners	Benefits	21%	0.44%
119	Cover all state/local government employees	Coverage	19%	0.32%
75	Reduce spouse benefit percentage	Family	17%	0.27%