JURNAL EMACS
[Engineering, MAthematics and Computer Science] Vol.7 No.2 May 2025: 191-200
e-ISSN: 2686-2573
DOI: 10.21512/emacsjournal.v7i2.13419

Adaptive Fuel Subsidy Optimization Using Deep Q-Learning and Bandit-Based Policy Selection: A Simulation Study

Pandu Dwi Luhur Pambudi
Computer Science Department, BINUS Online Learning,
Bina Nusantara University
Jakarta, Indonesia 11480
pandu.pambudi001@binus.ac.id

Correspondence: pandu.pambudi001@binus.ac.id

Abstract – *Designing effective fuel subsidy policies is a major challenge for governments seeking to balance energy affordability, fiscal sustainability, and environmental goals. This study introduces an adaptive simulation framework combining Deep Q-Learning and a multi-armed bandit algorithm to model fuel consumption behavior and optimize subsidy distribution strategies. Moreover, this paper simulates a dual-agent system in which a DQN-based consumer interacts with a bandit driven government selecting among three subsidy policies: universal, quota-based, and targeted. By simulating consumer responses to universal, quota-based, and targeted subsidies over 1,000 episodes, the framework demonstrates how policy can adapt in real-time to maximize social welfare and reduce inefficient spending. Results show that while universal subsidies often deliver the highest consumer satisfaction, they incur significant fiscal costs, whereas quota and targeted approaches can yield more balanced trade-offs. The study highlights the potential of reinforcement learning to enhance adaptive policymaking in complex economic systems. To strengthen the analysis, the simulation tracks both consumer and government rewards across scenarios, capturing the trade-off between satisfaction and fiscal burden. Evaluation results reveal that targeted subsidies, though less popular, often provide more sustainable outcomes. The agent-based approach enables the system to dynamically adjust policy decisions based on real-time feedback, reflecting the evolving nature of economic behavior. These findings underscore the usefulness of AI-driven simulations as decision-support tools in designing responsive and cost-efficient public policies.*

Keywords: *Reinforcement Learning; Bandit Algorithm; Fuel Subsidy; Policy Simulation; Q-learning*

I. INTRODUCTION

Fuel subsidies remain a contentious yet pivotal tool in public policy, aimed at mitigating fuel price volatility and supporting lower-income households. However, globally, fossil fuel subsidies exceeded $7 trillion in 2022 (Black et al., 2023), raising concerns about economic inefficiencies, regressive distributional effects, and environmental sustainability (International Monetary Fund, 2023). Universal subsidies, in particular, are widely criticized for disproportionately benefiting wealthier populations and distorting market incentives (Coady et al., 2015).

To address these challenges, many countries have piloted or implemented more nuanced mechanisms. Indonesia and Iran, for instance, have experimented with quota-based and targeted subsidies to encourage conservation and better reach vulnerable groups. This paper draws from such real-world practices by analyzing three representative subsidy schemes:

- Universal Subsidy: A flat-rate discount offered to all, regardless of income or usage level.
- Quota-Based Subsidy: A capped subsidy only available up to a threshold of consumption, promoting frugality.
- Targeted Subsidy: A means-tested benefit aimed at supporting low-income users.

Policymakers are increasingly exploring adaptive subsidy schemes that dynamically adjust to socioeconomic and behavioral factors. This adaptive potential aligns closely with recent advances in artificial intelligence particularly reinforcement learning (RL) -which enables agents to learn optimal policies from interactions with their environment (Kalatzantonakis et al., 2023; Lee et al., 2024; Li & Yu, 2021). Additionally, multi-armed bandit (MAB) algorithms provide scalable strategies for optimizing decisions under uncertainty (Zhang et al., 2024).

Energy costs are a key factor determining the success of businesses, particularly in energy-intensive industries. An analysis of revenue shares revealed that an increase in energy cost had a significant influence on the comparative net income

(Herman et al., 2023). Therefore, energy-related costs should be properly managed, and recent studies show that such costs can be increasingly optimized through intelligent, machine learning-based frameworks (Durairaj et al., 2022). In addition, firms adopting renewable energy sources frequently experience cost reductions through enhanced resource efficiency and diminished waste, promoting reputation and regulatory compliance while improving financial health (Hulshof & Mulder, 2020).

The global adoption of digital transformation has reshaped Fintech, notably in information products. As emphasized in Wu and Pambudi (2025), Fintech software manufacturers offer flexible pricing and security measures to accommodate new business models, such as SaaS subscriptions or one-time purchases. This move emphasizes options pricing as a strategic tool for managing security, scalability, and customer value—not merely a revenue function. The authors show how pricing strategies like on-premises premium and SaaS subscription pricing impact market position and profitability. Furthermore, the interaction between existing and new vendors for original and new demand reveals how entering firms can use pricing flexibility to gain market share, while existing firms prioritize maintaining profitability through innovations and loyalty policies. Related work by Wu and Pambudi (2024) examines vendor behavior in a competitive two-stage setting, showing that product bundling and managing customer churn significantly influence profitability. Bundling enhances perceived value and customer retention, supporting cooperative outcomes in competitive markets. Similarly, Siavvas et al. (2020) emphasize the role of security services in mitigating network effects, where strategic security investments can boost profit margins even without directly countering competitors. However, this effect weakens when rivals offer significantly superior products. These studies all use modeling and simulation techniques to assess policies and decisions related to pricing, product design, and security in competitive environments, highlighting the critical role these strategies play in shaping profitability under digital transformation and market competition (Saeed et al., 2024; Ogunleye et al., 2024).

Additionally, the application of machine learning models like Random Forest has gained significant traction in areas such as credit risk management (Kuyoro et al., 2022), fraud detection (Liu et al., 2015), and energy intensity analysis (Sahu & Pradhan, 2024), where such techniques are used to improve accuracy, feature relevance, and decision-making. These predictive models play a crucial role in enhancing energy efficiency while simultaneously optimizing operational strategies (Nadkarni et al., 2023; Rubio et al., 2021). Furthermore, machine learning techniques such as Random Forest are favored over other models like Gradient Boosting Machines for their interpretability and robustness to noise (Nadkarni et al., 2023). In financial analysis, these models are leveraged to assess energy consumption and security investments, where the combined effects significantly influence profitability (Siavvas et al., 2020). These prior works indicate the widespread use of modeling and simulation to assess optimal decisions across various domains, from pricing strategies to security and energy efficiency (Saeed et al., 2024; Ogunleye et al., 2024; Wu and Pambudi, 2023).

While RL and MAB techniques have been applied in areas such as energy efficiency (Cunha et al., 2022), finance (Ni et al., 2023), and healthcare (Zhang et al., 2024), their use in simulating and optimizing adaptive fuel subsidies remains underexplored. This study addresses that gap by modeling a two-agent system: a DQN-based consumer interacting with a bandit-driven government subsidy policy selector. The simulation aims to know trade-offs between efficiency and equity across subsidy types and highlights the potential of machine learning for real-time economic policy design. The simulation reveals that while universal subsidies deliver consistent consumer rewards, they result in persistently negative government returns-highlighting a critical trade-off between political popularity and fiscal sustainability. The results demonstrate the viability of integrating AI into dynamic public finance simulations and inform future directions for policy reform.

II. METHODS

2.1. Environment Design

The simulation environment is constructed as a stylized representation of a simplified economic interaction between a single consumer and a central government that implements fuel subsidy policies. The environment is framed as a discrete-time Markov Decision Process (MDP), where transitions occur across time steps t, driven by agent actions and stochastic state evolutions.

The state at each time step is encoded as a continuous-valued vector $\mathbf{s}_t \in \mathbb{R}^3$, composed of three normalized features: consumer income (s_t^1), previous fuel demand (s_t^2), and a fixed base fuel price (s_t^3). These inputs capture both the economic capability of the consumer and the inertia of recent consumption behavior. All values are scaled to lie within the interval $[0,1]$ for compatibility with neural network-based agents and to ensure generality across diverse economic contexts.

The action space $\mathbf{a}_t \in \{0,1,\dots,9\}$ represents discrete choices available to the consumer, where each action maps linearly to a fuel demand level

ranging from 0 to a maximum quota of 50 units. Thus, action 0 corresponds to 0 units of consumption, while action 9 corresponds to the full quota. This discrete formulation simplifies learning and aligns with realistic policy levers such as rationed access to subsidized fuel.

The government, acting as a policy setter, chooses one of three predefined subsidy mechanisms to influence the consumer's effective fuel price:

- Universal Subsidy: Applies a flat discount of 0.3 to all consumers, irrespective of income or consumption behavior.
- Quota-Based Subsidy: Grants a discount of 0.2 only when the fuel demand is less than or equal to 25 units, promoting conservation.
- Targeted Subsidy: Allocates a larger discount of 0.4 for consumers with normalized income levels below 0.5, simulating needs-based assistance.

The effective price per unit of fuel, after policy intervention, is then computed as:

$$P_{\text{eff}} = \max\left(0.1, P - \delta_{\text{policy}}\right) \quad (1)$$

where P is the fixed base fuel price (set to 0.5 in the simulation) and δ_{policy} is the subsidy amount determined by the active policy. A lower bound of 0.1 prevents unrealistically low prices and maintains economic plausibility.

The consumer's reward R_C reflects net utility, formulated as the product of income and demand minus the total fuel expenditure:

$$R_C = \text{income} \times \text{demand} - \text{demand} \times P_{\text{eff}} \quad (2)$$

This structure incentivizes higher utility through fuel access while penalizing excessive expenditure. In contrast, the government's reward R_G is designed to incorporate both the consumer's utility (reflecting welfare) and the fiscal cost of the subsidy:

$$R_G = R_C - \left(\text{demand} \times \delta_{\text{policy}}\right) \quad (3)$$

This dual-objective function captures the trade-off faced by policymakers: maximizing consumer welfare while minimizing subsidy spending. Through repeated interaction between the consumer and the policy environment, this setup enables the evaluation of subsidy efficiency and behavioral adaptation over time.

Overall, the environment is tailored to study adaptive economic behavior under constrained resources, and to explore how intelligent agents can learn or evolve policies that optimize long-term welfare in socioeconomically diverse contexts.

2.2. Consumer Agent: Deep Q-Network

The consumer agent in this simulation framework is modeled using a Deep Q-Network (DQN), a value-based reinforcement learning algorithm that approximates the optimal action-value function $Q(s, a)$. This function estimates the expected cumulative reward for taking action a in state s and thereafter following the optimal policy.

The learning process is governed by the Bellman optimality equation, which provides a recursive formulation for updating the Q -values:

$$Q(s_t, a_t) = \mathbb{E}\left[r_t + \gamma \max_{a'} Q(s_{t+1}, a')\right] \quad (4)$$

Here, r_t denotes the immediate reward received at time $t, \gamma \in [0,1)$ is the discount factor reflecting the relative importance of future rewards, and $\max_{a'} Q(s_{t+1}, a')$ represents the maximum expected reward achievable from the next state s_{t+1}.

The Q-function is approximated using a deep neural network with multiple hidden layers, capable of capturing complex, non-linear relationships between the input state space and the resulting Q-values. The state vector s_t comprises normalized economic indicators including consumer income, previous fuel demand, and market fuel price. The discrete action space represents the consumer's choice of fuel quantity, scaled to reflect a maximum quota.

Training of the DQN is implemented using the stable-baselines3 library and spans 10,000 timesteps. The algorithm employs experience replay, a technique wherein past transitions (s_t, a_t, r_t, s_{t+1}) are stored in a memory buffer and randomly sampled in batches for network training. This reduces the temporal correlations in the training data and improves sample efficiency.

To balance exploration and exploitation, the agent follows an ε-greedy policy. With probability ε, the agent selects a random action to explore new behaviors, while with probability $1 - \varepsilon$, it chooses the action that maximizes the current Q-value estimate. This mechanism ensures that the agent continues to explore suboptimal policies and avoids getting trapped in local optima during early training phases.

The combination of function approximation, temporal-difference learning, and randomized exploration equips the consumer agent with the capability to learn optimal consumption behaviors under varying policy regimes and economic scenarios. Over the course of training, the agent converges toward a strategy that maximizes individual utility in response to dynamically changing fuel subsidy policies.

2.3. Government Agent: Upper Confidence Bound Bandit

The government agent is tasked with selecting an optimal fuel subsidy policy from a discrete set of three options: universal, quota-based, and targeted. To manage this decision-making process in a data-driven and adaptive manner, the agent utilizes a

multi-armed bandit (MAB) framework-specifically, the Upper Confidence Bound (UCB) algorithm.

The UCB algorithm is designed to balance the trade-off between exploration (trying less-known policies to gather information) and exploitation (choosing policies that have performed well historically). For each policy $\mathrm{arm}i$, the agent maintains a cumulative reward estimate $\hat{\mu}_i$ and a selection count n_i. The UCB value at time step t for each arm is computed as follows:

$$\mathrm{UCB}_i(t) = \hat{\mu}_i + \sqrt{\frac{2\log\ t}{n_i}} \tag{5}$$

Here, the first term represents the empirical mean reward, while the second term is an exploration bonus that diminishes as n_i increases, encouraging the agent to try policies with lower selection frequencies.

After each episode, the reward received by the government agent, denoted R_G, is used to update the estimate $\hat{\mu}_i$ of the selected policy i via incremental averaging:

$$\hat{\mu}_i \leftarrow \hat{\mu}_i + \frac{1}{n_i}(R_G - \hat{\mu}_i) \tag{6}$$

This allows the agent to refine its estimate based on new observations while maintaining computational efficiency and numerical stability. To ensure sufficient initial exploration of all policy options, an ε-greedy mechanism is incorporated with $\varepsilon = 0.1$. This means that 10% of the time, the agent will select a policy at random, regardless of its UCB score, thereby reducing the risk of premature convergence to suboptimal policies.

This hybrid exploration strategy - combining the statistical rigor of UCB with random sampling via ε -greedy-enhances the agent's ability to discover the most fiscally efficient and socially beneficial policy over time. The design reflects real-world policy experimentation where a balance is often sought between established programs and innovative interventions.

2.4. Simulation Protocol

To evaluate the adaptive dynamics between subsidy strategies and consumption behavior, the simulation was executed over 1,000 discrete episodes. In each episode, a sequence of interactions unfolds between three key components: the simulation environment, a consumer agent governed by a Deep Q-Network (DQN), and a government agent implementing a multi-armed bandit policy selection strategy.

At the beginning of each episode, the environment is initialized with a randomly sampled consumer income and demand history, thereby ensuring heterogeneity in agent experiences. The government agent selects one of three available subsidy policies-universal, quota-based, or targeted-using the Upper Confidence Bound (UCB) algorithm, which weighs both the historical effectiveness of each policy and the uncertainty associated with underexplored options. This selection reflects a balance between exploiting high-performing policies and exploring less-tested strategies to improve long-term performance.

The chosen policy is then passed to the consumer agent, which observes the current economic state vector and selects a fuel consumption level using its trained Q-network. The environment calculates both consumer utility and government fiscal impact based on the interaction between the selected policy and the agent's demand response. Specifically, the simulation captures both the direct utility gained by the consumer and the cost burden imposed on the state due to the subsidy expenditure.

Reward values for both agents are computed and logged at each step, with a focus on tracking the evolution of policy efficiency over time. To mitigate noise in the reward trajectory and enhance interpretability, a rolling average with a window size of 50 episodes is applied. This smoothing technique allows for the visualization of performance trends, particularly in assessing convergence behavior and the stability of policy preferences.

Overall, the simulation protocol is designed not only to assess the static performance of individual policies but also to observe how adaptive learning algorithms respond under dynamic and uncertain conditions. The emergent behavior from these agent-policy interactions provides insights into the robustness and practical viability of AIassisted economic policy design.

2.5. Code Implementation

To operationalize the described model, the simulation was implemented using Python with support from key libraries such as gymnasium, numpy, pandas, matplotlib, and stable-baselines3. The core logic was structured around a custom reinforcement learning environment compliant with the OpenAI Gym interface, allowing seamless integration with RL algorithms.

The consumer agent was modeled using the DQN algorithm provided by stable-baselines3, and trained over 10,000 timesteps. The government agent was implemented as a multi-armed bandit using the Upper Confidence Bound (UCB) strategy, supported by a greedy mechanism to encourage initial exploration.

The following Python code snippet summarizes the key elements of the implementation:

```
--- ENVIRONMENT ---
class FuelConsumptionEnv(gym.Env):
def init(self):
... # Environment variables and
limits
def reset(self):
... # State initialization
def step(self, action, policy=None):
... # Reward calculation and state
update
```

```
--- DQN AGENT (Consumer) ---
env = make_vec_env(lambda:
FuelConsumptionEnv(), n_envs=1)
model = DQN("MlpPolicy", env,
verbose=0)
model.learn(total_timesteps=10000)
```

```
--- BANDIT AGENT (Government) ---
class GovernmentBanditAgent:
def init(self):
... # UCB setup and reward tracking
def select_arm(self, epsilon=0.1):
... # UCB logic with epsilon-greedy
exploration
def update(self, arm, reward):
... # Reward update rule
--- SIMULATION LOOP ---
for t in range(1000):
... # Agent interaction, reward
logging, plotting
```

2.6. Simulation Workflow Overview

Figure 1 presents the process flow diagram outlining the simulation logic behind the adaptive fuel subsidy framework. This schematic illustrates the sequence of operations performed during each episode of the 1,000 -step simulation, highlighting the interplay between the consumer agent, the government agent, and the environment.

The workflow begins at the Start node, followed by the Initialization of the Environment, where a new instance of the simulation is created. This includes random sampling of consumer-specific variables such as income and initial demand, ensuring variability across episodes and simulating a dynamic economic environment.

Next, in the Initialize Agents phase, two intelligent agents are activated:

- A government agent employing an Upper Confidence Bound (UCB) algorithm enhanced with an ε-greedy exploration strategy for adaptive policy selection.
- A consumer agent modeled using Deep Q-Learning (DQN), trained to optimize fuel consumption decisions under varying economic states and subsidy schemes.

The process then bifurcates into two parallel computations:

- The government selects one of three available subsidy policies: universal, quota-based, or targeted.
- Simultaneously, the DQN-based consumer predicts the optimal fuel consumption action based on its learned Q-value function and current environmental state.

Both outputs converge in the Step Environment node, where the simulation environment computes the immediate consequences of the consumer's action under the selected policy. The new state is generated, and reward signals for both agents are calculated.

Subsequently, in the Compute Rewards phase, two distinct reward functions are evaluated:

- Consumer reward represents net utility, calculated as the difference between consumption value and cost.
- Government reward incorporates fiscal expenditure on subsidies, reflecting the trade-off between policy generosity and budgetary efficiency.

Following the reward evaluation:

- The government agent updates its UCB estimates based on received feedback, refining its internal value representation of each policy.
- Simultaneously, reward and policy logs are maintained for analysis, allowing performance tracking over time.

The Plot Results stage visualizes two outputs: smoothed consumer and government reward trajectories, and the frequency distribution of policy selections. These outputs offer valuable insights into policy performance and behavioral adaptation.

Finally, the process concludes at the End node, completing a single iteration. This structure is repeated for 1,000 episodes to assess long-run convergence, learning dynamics, and policy implications. This modular simulation structure facilitates extensibility for more complex scenarios, including multi-agent setups, contextual bandit formulations, or the integration of broader economic indicators.

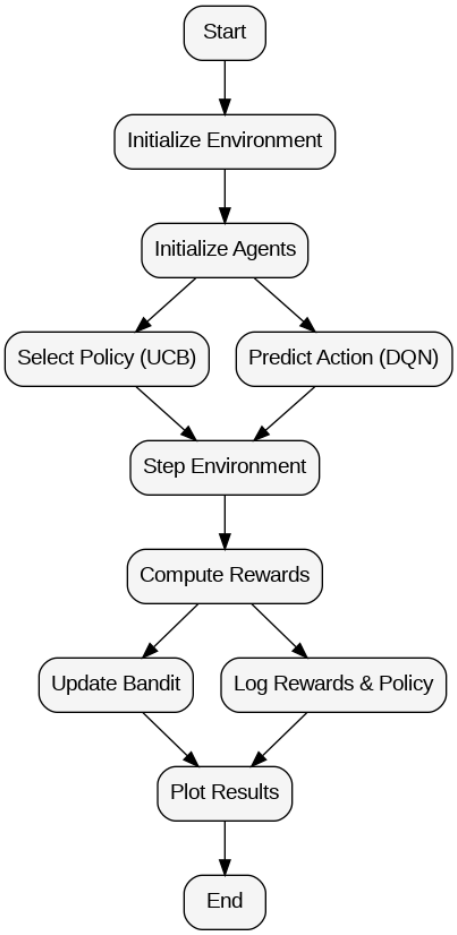

Figure 1. Simulation flowchart for adaptive fuel subsidy evaluation using Deep Q-Learning and UCB bandit agents.

2.7. Parameter Configuration Overview

To ensure reproducibility and clarity, Table 1 summarizes the key parameters used in the simulation, including environment setup, agent configurations, and algorithmic hyperparameters.

This parameter summary serves as a reference point for interpreting the simulation results and supports reproducibility for future studies extending this framework.

Table 1. Simulation and Agent Parameters

Category	Parameter	Description
Environment	Max quota	50 units - maximum possible fuel demand
	Price floor	0.1 - prevents unrealistically low fuel prices
	Base fuel price	0.5 - fixed market price before subsidy
	State vector	[income, previous demand, price]
Consumer Agent (DQN)	Algorithm	Deep Q-Network (DQN)
	Library	stable-baselines3
	Total timesteps	10,000 - training length
	Exploration strategy	ε-greedy, $\varepsilon = 0.1$
	Experience replay	Enabled - improves sample efficiency
Government Agent (Bandit) $\boldsymbol{\varepsilon}$	Policy options	[universal, quota, targeted]
	Exploration strategy	UCB with ε-greedy
		0.1 - random exploration probability
	UCB formula	
	Reward update	Incremental average update of $\hat{\mu}_i$
Simulation Protocol	Iterations	1,000 episodes
	Smoothing window	50 - for reward visualization

III. RESULTS AND DISCUSSION

The results show that Figure 2 presents smoothed consumer and government rewards over 1,000 simulation steps. Consumer rewards remained consistently positive, indicating effective adaptation to varying policy environments. In contrast, government rewards fluctuated and were often negative, reflecting the high fiscal cost associated with certain.

Figure 2. Smoothed Reward over Time (window=50)

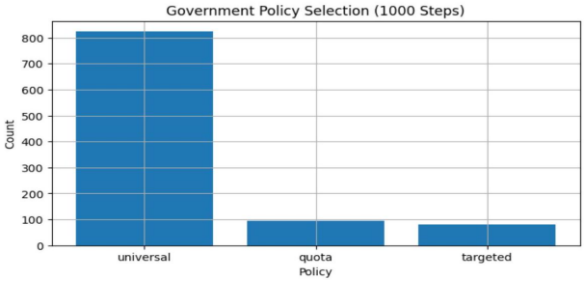

Figure 3. Government Policy Selection (1000 Steps) subsidy strategies

Figure 3 illustrates the frequency of policy selection by the government agent. Universal subsidies were selected in over 80% of episodes, highlighting their strong consumer performance but weak government efficiency.

This underscores the classic tension between popular and sustainable policy choices. The limited exploration of quota and targeted subsidies may have led to an underestimation of their long-term potential, pointing to a key weakness in the current bandit strategy. This aligns with findings from (Thibodeau et al., 2024), who emphasize the importance of balancing exploration and exploitation in policy selection to avoid suboptimal long-term outcomes.

In the evaluation matrix, the simulation results show an average consumer reward of 9.28, indicating that the consumers generally benefited from the subsidy policies, especially from the universal and targeted subsidies. This positive reward reflects the consumer's satisfaction, as the reward is directly linked to their utility, which depends on factors like income and fuel demand. However, the government's average reward is negative at -0.44, suggesting that the subsidies, particularly the universal subsidy, result in significant fiscal costs, making the policies less sustainable from a financial perspective.

The policy selection frequency highlights the preferences of the government agent, with the targeted subsidy being chosen most often (720 times), followed by the universal subsidy (196 times), and the quota-based subsidy (84 times). This preference suggests that the government favors policies that provide support to low-income consumers while minimizing fiscal burden. The universal subsidy, despite its popularity with consumers, is selected less frequently due to its higher cost to the government, which may not be sustainable in the long term.

The ANOVA test further supports the validity of these observations, with a statistically significant difference (F-statistic: 13.06, p-value: 0.0000) between the rewards of the three policies. This result confirms that the rewards from the three subsidy policies are not equal, emphasizing the varying effectiveness and fiscal impact of each policy. The significant p-value indicates that the differences in consumer rewards across the policies are not due to random chance, reinforcing the importance of choosing the right subsidy strategy based on the desired outcomes for both consumers and the government.

Therefore, this paper provides four managerial insights. (1) *Balancing Popularity and Fiscal Sustainability***.** The simulation demonstrates the allure of universal subsidies due to their broad appeal and stable consumer outcomes. However, this strategy places a substantial burden on government budgets. This trade-off mirrors real-world dilemmas, such as in Malaysia, where the government is preparing to reduce fuel subsidies in mid-2025 to address fiscal constraints Bloomberg (2024). (2) *Importance of Adaptive Policy Mechanisms***.** The dynamic interactions modeled in the simulation highlight the need for responsive, data-driven policy tools. Adaptive policies that adjust based on real-time feedback can prevent wasteful spending. Indonesia's ongoing evaluation of its fuel subsidy scheme, aiming for a potential phase-out by 2027, illustrates a national commitment to reform (Reuters, 2024). (3) *Targeted Subsidies as a Viable Alternative.* Despite being underutilized in the simulation, quota and targeted strategies hold promise for efficiency and equity. India's recent decision to increase targeted subsidies (e.g., cooking gas, fertilizer) by 8% in its 2025 budget suggests a growing preference for more focused interventions (Reuters, 2025). (4) *Leveraging Technology for Policy Optimization.* The successful integration of reinforcement learning and multi-armed bandit algorithms in this study supports the growing role of AI in policy design, as highlighted in prior research (Mui et al., 2021; Oda et al., 2022; Xu et al., 2020). These methods enable simulations to reveal effective long-term strategies, offering a scientific basis for navigating complex fiscal decisions (Thibodeau et al., 2024).

In conclusion, this analysis reveals that while universal subsidies may be politically expedient, they are not always economically optimal. Future subsidy reforms should incorporate adaptive mechanisms, prioritize targeted support, and consider leveraging AI to navigate the tension between equity and fiscal responsibility.

IV. CONCLUSION

This study provides a novel contribution to the intersection of artificial intelligence and public economic policy by demonstrating the potential of combining deep reinforcement learning (DRL) and bandit algorithms to inform adaptive fuel subsidy strategies. By simulating an interactive environment between a consumer agent-trained through Deep Q-Learning - and a government agent-guided by a multi-armed bandit decision model-we explore how varying subsidy policies can evolve dynamically based on feedback from consumption behavior and policy performance.

The results indicate that the consumer agent is capable of learning optimal consumption behaviors under a range of economic conditions, validating the applicability of reinforcement learning in modeling realistic and adaptive decision-making. Moreover, the government agent consistently favored the universal subsidy policy due to its relatively higher short-term reward from the consumer's perspective. However, this preference came at a fiscal cost, resulting in suboptimal outcomes for government welfare. This divergence between the objectives of

consumer satisfaction and governmental efficiency reflects a common policy dilemma, emphasizing the need for models that can better balance equity and sustainability.

The broader implications of this work lie in its support for the integration of AI into policy simulation and economic modeling. The use of reinforcement learning and bandit strategies offers a powerful framework for designing public policies that are responsive to changing conditions and capable of optimizing complex tradeoffs. However, this approach also highlights the challenge of aligning technological optimization with normative public values, such as fairness and long-term social equity.

Despite its contributions, the study is not without limitations. The simulation environment is intentionally simplified to allow for computational feasibility and clarity of interpretation, which necessarily limits the scope of realism. The single-agent setup does not capture heterogeneity among consumers, such as differences in income, geography, or behavioral tendencies, which are critical in real-world policy design. Furthermore, the government agent lacks contextual awareness, relying solely on aggregated reward signals without considering macroeconomic trends or social priorities. The reward structures themselves are also linear and may not fully reflect the multi-dimensional objectives of public policy.

Looking ahead, future research should expand the framework to include multi-agent architectures that can simulate a diverse population of consumers with varying economic and behavioral profiles. Introducing contextual bandit models would also allow the government agent to tailor subsidy strategies to specific demographic or economic segments. Additionally, the integration of broader evaluation metrics - such as income inequality, carbon emissions, or long-term budgetary impacts- would enhance the model's utility for real-world policymaking. These extensions would move the framework closer to the complexity of actual subsidy systems and increase its relevance for policy experimentation and reform in diverse national contexts.

ACKNOWLEDGMENT

The author would like to express their sincere gratitude to Bina Nusantara University for providing the support necessary for the publication of this research. Their continuous encouragement and resources have been invaluable in the completion of this study. I deeply appreciate their commitment to fostering academic research and innovation.

AUTHOR CONTRIBUTIONSHIP

Pandu Dwi Luhur Pambudi: A Correspondent Author, Conceptualization, Methodology, Resources, Formal Analysis, Writing – Original Draft Preparation, Data Simulation and Curation, Investigation, Validation, Analysis.

DECLARATION OF COMPETING INTEREST

No author associated with this paper have disclosed any potential or pertinent conflicts and have no known competing financial interests that may be perceived to have an impending conflict with this work.

DISCLOSURE INSTRUCTIONS

During the preparation of this work the author used ChatGPT 4.0 in order to improve the flow of the text. After using this tool/ service, the author reviewed and edited the content as needed and take full responsibility for the content of the publication.

DATA AVAILABILITY

To support further research, I have made the data and simulation used in this study publicly accessible. The simulation and data can be accessed via the following links: *https://github.com/pandu1992/PPB_Research/blob/main/SINTA_adaptive_fuel_subsidy_simulation_fixed.ipynb*.

REFERENCES

Black, S., Parry, I., & Vernon, N. (2023). Fossil fuel subsidies surged to record $7 trillion. IMF blog, 24. Retrieved from https://www.imf.org/en/Publications/WP/Issues/2023/08/22/IMF-Fossil-Fuel-Subsidies-Data-2023-Update-537281

Bloomberg. (2024). Malaysia minister braces for backlash over fuel subsidy revamp. Retrieved from https://www.bloomberg.com/news/articles/2024-10-20/malaysia-minister-braces-for-backlash-over-fuel-subsidy-revamp?embedded-checkout=true

Coady, M. D., Parry, I. W., Sears, L., & Shang, B. (2015). How large are global energy subsidies? International Monetary Fund.

Cunha, R. F., Gonçalves, T. R., Varma, V. S., Elayoubi, S. E., & Cao, M. (2022). Reducing fuel consumption in platooning systems through reinforcement learning. IFAC-PapersOnLine, 55(15), 99-104.

Durairaj, D., Wroblewski, L., Sheela, A., Hariharasudan, A., & Urbanski, M. (2022).

Random forest based power sustainability and cost optimization in smart grid. Production Engineering Archives, 28(1), 82–92.

Herman, R., Nistor, C., & Jula, N. M. (2023). The influence of the increase in energy prices on the profitability of companies in the European Union. Sustainability, 15(21), 15404.

Hulshof, D., & Mulder, M. (2020). The impact of renewable energy use on firm profit. Energy Economics, 92, 104957.

Kalatzantonakis, P., Sifaleras, A., & Samaras, N. (2023). A reinforcement learning-variable neighborhood search method for the capacitated vehicle routing problem. Expert Systems with Applications, 213, 118812.

Kuyoro, A. O., Ogunyolu, O. A., Ayanwola, T. G., & Ayankoya, F. Y. (2022). Dynamic effectiveness of random forest algorithm in financial credit risk management for improving output accuracy and loan classification prediction. Ingenierie des systèmes d'information, 27(5), 815–821.

Lee, S., Liebana, S., Clopath, C., & Dabney, W. (2024). Lifelong reinforcement learning via neuromodulation. arXiv preprint arXiv:2408.08446.

Li, J., & Yu, T. (2021). Optimal adaptive control for solid oxide fuel cell with operating constraints via large-scale deep reinforcement learning. Control Engineering Practice, 117, 104951.

Liu, C., Chan, Y., Kazmi, S. H. A., & Fu, H. (2015). Financial fraud detection model: Based on random forest. International Journal of Economics and Finance, 7(7), 27–35.

Mui, J., Lin, F., & Dewan, M. A. A. (2021). Multi-armed bandit algorithms for adaptive learning: A survey. In International Conference on Artificial Intelligence in Education (pp. 273-278). Cham: Springer International Publishing.

Ni, H., Xu, H., Ma, D., & Fan, J. (2023). Contextual combinatorial bandit on portfolio management. Expert Systems with Applications, 221, 119677.

Nadkarni, S. B., Vijay, G., & Kamath, R. C. (2023). Comparative study of random forest and gradient boosting algorithms to predict airfoil self-noise. Engineering Proceedings, 59(1), 24.

Oda, A., Mihana, T., Kanno, K., Naruse, M., & Uchida, A. (2022). Adaptive decision making using a chaotic semiconductor laser for multi-armed bandit problem with time-varying hit probabilities. Nonlinear Theory and Its Applications, IEICE, 13(1), 112-122.

Ogunleye, O., Adeniji, S., Onih, V., Simo, Y., Elom, E., Kanu, E., ... & Ejiofor, O. (2024). Improving resilience and efficiency in the energy sector: A perspective on cybersecurity and renewable energy storage. Valley International Journal Digital Library, 502-513.

Reuters. (2024). Indonesia weighs plan to phase out fuel subsidies by 2027. Retrieved from https://www.reuters.com/business/energy/indonesia-conducting-thorough-exercise-reform-fuel-subsidy-scheme-minister-says-2024-11-04/

Reuters. (2025). India's budget likely to raise major subsidies by 8% to $47 billion in next fiscal. Retrieved from https://www.reuters.com/world/india/indias-budget-likely-raise-major-subsidies-by-8-47-bln-next-fiscal-2025-01-22/

Rubio, F., Llopis-Albert, C., & Valero, F. (2021). Multi-objective optimization of costs and energy efficiency associated with autonomous industrial processes for sustainable growth. Technological Forecasting and Social Change, 173.

Saeed, S., Gull, H., Aldossary, M. M., Altamimi, A. F., Alshahrani, M. S., Saqib, M., ... & Almuhaideb, A. M. (2024). Digital transformation in energy sector: Cybersecurity challenges and implications. Information, 15(12), 764.

Sahu, U. K., & Pradhan, A. K. (2024). Discovering the determinants of energy intensity of Indian manufacturing firms: A panel data approach. Discover Sustainability, 5(1), 139.

Siavvas, M., Marantos, C., Papadopoulos, L., Kehagias, D., Soudris, D., & Tzovaras, D. (2020). On the relationship between software security and energy consumption. IEEE Transactions on Emerging Topics in Computing, 8(3), 535–545.

Thibodeau, J., Nekoei, H., Taïk, A., Rajendran, J., & Farnadi, G. (2024). Fairness incentives in response to unfair dynamic pricing. arXiv preprint arXiv:2404.14620.

Wu, C. H., & Pambudi, P. D. L. (2025). Digital transformation in fintech: Choosing between application and Software as a Service (SaaS). Asia Pacific Management Review, 30(2), 100342.

Wu, C. H., & Pambudi, P. D. L. (2024). On-premise software vs. cloud-based software under the presence of product bundling. Procedia Computer Science.

Wu, C. H., & Pambudi, P. D. L. (2023). Exploring security service on information product's pricing decisions. In 2023 5th International Conference on Management Science and Industrial Engineering (pp. 159-163). New York, NY, USA: ACM.

Xu, S., Wang, F., Wang, H., & Romberg, J. (2020). In-field performance optimization for mm-

wave mixed-signal Doherty power amplifiers: A bandit approach. IEEE Transactions on Circuits and Systems I: Regular Papers, 67(12), 5302-5315.

Zhang, K. W., Closser, N., Trella, A. L., & Murphy, S. A. (2024). Replicable bandits for digital health interventions. arXiv preprint arXiv:2407.15377.