INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Investment Strategy for Informational Assets in Oil and Gas Exploration Using Deep Reinforcement Learning

Paulo Roberto de Melo Barros Júnior

Master in Business Economics and Competition

Supervisor:
PhD Mónica Meireles, Assistant Professor,
ISCTE Business School

September, 2024

BUSINESS
SCHOOL

Department of Economics

Investment Strategy for Informational Assets in Oil and Gas Exploration Using Deep Reinforcement Learning

Paulo Roberto de Melo Barros Júnior

Master in Business Economics and Competition

Supervisor:
PhD Mónica Meireles, Assistant Professor,
ISCTE Business School

September, 2024

This project is dedicated to my family, friends, teachers, and coworkers in Petrobras, who encourage me to keep moving in a continuous learning pathway.

Acknowledgment

First, I express my deepest gratitude to Petrobras for financing this master's program and sponsoring me. Their support, along with the infrastructure and resources provided by ISCTE, has been invaluable in developing this project. I sincerely thank my colleagues at ISCTE, Paulo Quimbango and family, Carmo Fernandes, and Rui Batatinha. Your collaboration and support have been crucial throughout this journey.

A special acknowledgment goes to my supervisor, Mónica Meireles, my teacher, Nádia Simões, and my Co-Supervisor, José Luis Silva, for their invaluable knowledge and availability during the thesis development. Your guidance has been instrumental in my academic growth.

I am profoundly grateful to the managers at Petrobras Guilherme Canha, my career guide and leader, and to Jeferson Kinzel and Jonilton Pessoa for their confidence and the opportunities they provided. To my partners at Petrobras, Sandinei, for showing me this path, and Nicolai, for managing things at work, your support has been essential. I also thank all my colleagues in the department who have taught me so much and shared many good times at work. I must thank Catiana, Nunes, and Morgana for being our partners in Portugal and for being the wonderful people we were lucky to meet.

Lastly, I would like to extend a special acknowledgment to my family, especially mother Leni and siblings Daniel and Gabi, and their aggregates Pedro Dan, Patricia e Eduardo, and my aunt Celi, for everything they have done for me. To my beloved companions, Fabi, Bibi, and Mag, your unwavering support and companionship have been my most significant source of strength and joy. I am deeply grateful for your love and presence in my life and genuinely thankful for your contributions.

Abstract

This thesis investigates the investment strategies in informational assets in the exploratory phase of the oil and gas industry. We apply a Reinforcement Learning (RL) framework to simulate economic scenarios with multiple agents, actions, and environments to identify optimal investment approaches. Our approach determines the most effective policies under different economic conditions by implementing Q-Learning, SARSA, and Deep Q-Network (DQN) algorithms. The results evidence the efficacy of RL-trained agents in attaining superior returns, particularly in competitive bidding scenarios where a smaller number of companies implies a higher probability of success and maximizes the benefits of advanced investments in informational assets. For scenario features, we observe that while oil prices and demand rise, the returns increase. However, we can not observe significant changes in the advantages of early investment in informational assets. The RL system and database developed in this study provide a foundation for real-world application in the oil and gas industry, with the potential for enhancements in modeling states, actions, and agents or the incorporation of advanced techniques such as A3C and DPO. We highlight the potential of RL in complex decision-making processes and deliver a robust tool for optimizing investment strategies. It also provides a valuable framework for use in business environments in the oil and gas industry or with similar characteristics.

Keywords: Oil & Gas Exploration, Investment Strategy, Reinforcement Learning.
JEL: D46, D81, G30.

Resumo

Nesta tese investigamos as estratégias de investimento em ativos informacionais na fase exploratória da indústria de petróleo e gás. Aplicamos uma estrutura de Aprendizagem por Reforço (AR) para simular cenários económicos com múltiplos agentes, ações e ambientes com o intuito de identificar estratégias de investimento ótimas. Esta abordagem determina as políticas mais eficazes sob diferentes condições económicas, implementando algoritmos Q-Learning, SARSA e Deep Q-Network (DQN). Os resultados evidenciam a eficácia dos agentes treinados em RL na obtenção de retornos superiores, particularmente em cenários de licitações competitivas onde um número menor de empresas implica uma maior probabilidade de sucesso e maximiza os benefícios de investimentos antecipados em ativos informacionais. Para as características relativas aos diferentes cenários económicos, observamos que enquanto os preços e a procura do petróleo aumentam, os retornos também aumentam. Contudo, não vemos mudanças significativas nas vantagens do investimento antecipado em ativos informacionais. O sistema AR e o banco de dados desenvolvidos neste estudo fornecem uma base para aplicação ao mundo real na indústria de petróleo e gás, com potencial para melhorias na modelização de estados, ações e agentes ou para a incorporação de técnicas avançadas como A3C e DPO. A investigação evidencia o potencial da AR em processos complexos de tomada de decisão, oferecendo uma ferramenta robusta para otimizar estratégias económicas e fornecendo uma estrutura valiosa para utilização em ambientes de negócios com atividades análogas às da indústria de petróleo e gás.

Palavras-chave: Exploração de Óleo & Gás, Estratégia de Investimento, Aprendizagem por Reforço.
JEL: D46, D81, G30.

Contents

List of Figures

List of Tables

Introduction

This master's thesis investigates solutions for a real challenge in the Oil & Gas Industry (O&GI). The goal is to optimize the decision-making process within the oil and gas industry for investments in information assets during the exploration phase. In our industry, a big issue is deciding how to invest money in high-quality but expensive information versus lower-quality but more affordable information. The oil and gas business is risky and unpredictable. In this situation, having good data that is available and accurate is really important for making smart decisions about both strategy and day-to-day operations. If a company has a good understanding of the geology of specific areas, especially insights into the unique characteristics of sedimentary basins, it can gain an edge by finding and securing promising oil production areas at a better price and with a more realistic bid than its competitors.

This knowledge helps companies accurately assess potential reserves and production capacity within a specific prospect. This allows them to allocate resources more efficiently, optimize their investments, and maximize returns. However, obtaining such valuable information assets comes with its own set of challenges. Acquiring high-quality data often requires a significant upfront capital outlay as early-stage investments in exploration and data acquisition technologies. Investing in information assets in the early stages usually involves taking a calculated risk [**1**].

If the data obtained does not end up enabling the successful identification, securing and development of a commercial oil field, then the upfront investments can turn into sunk costs. This may result in poor economic performance for the company and disturb future exploration initiatives. Therefore, attention must be given to potential benefits of improved data quality, as well as the risks of failing to do so. The key to digital transformation initiatives relies on understanding how and where data should be acquired, and what analytical methods would be more beneficial under certain circumstances. These scenarios might include the geological potential in a target area, competitive analysis or simply how much risk a company can afford. It might be essential to develop a plan B that reduces the risk when exploration does not produce reliable findings. Hence, investing in information assets could offer a tangible competitive edge, but these gains should be fully weighed against the associated risks [**2**].

The advancement of technology has positioned information as a crucial player for all industries across the economy. This stylized fact impacts the mining exploration industry even more since this sector requires crucial input data to model the geology under investigation. Our research is focused on helping decision-makers finding the best strategy for investing in information acquisition, considering a large group of parameters and scenarios related to the type of data, survey costs, and features of information, such as quality, perishability, and usability. To

illustrate, we could explore the set of technical parameters that can impact information costs, quality and temporal availability, as well as the environmental degradation associated with acquiring each category type. The potential effect of this research on the Oil & Gas Industry is noteworthy, providing optimism for a more efficient, robust, and effective decision-making process.

In agreement with [**3**], a typical oil exploration dataset comprises seismic volumes and well data. From this data it is possible to create a subsurface model structure with horizons and faults, as shown in Figure 1. To collect these kinds of data, service companies conduct a survey with large vessels operating in specific areas for periods ranging from three months to one year. A seismic or well crew team's daily cost varies from US$ 100k to US$ 500k. These figures provide an idea of the investment involved in oil exploration. For seismic data alone, the global maritime seismic acquisition industry invests, on average, over US$ 10 billion per year, with projected costs reaching US$ 20 billion depending on the economic scenario[**4**].

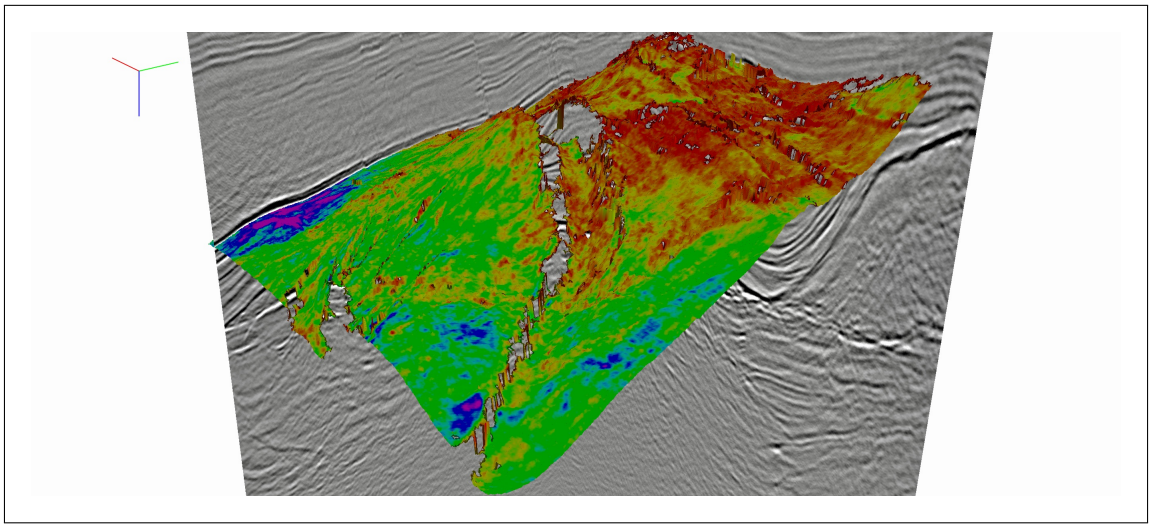
To simulate the current and alternative strategies for investing in information assets, we utilize advanced machine learning (ML) techniques. First, we focus on understanding the factors guiding decision makers from O&G companies. Second, we investigate and point out the optimal allocation investments for various economic scenarios.

The development of the model allows us to analyze optimal strategies based on the return across the entire value chain of the oil and gas industry. We consider the return in each phase, bid, exploration, development, and production, as well as the quality of information and investment applied for a set of firms representing the offshore upstream market investors. These two variables are converted into actions and states for agents that learn how to decide on the necessary quality to effectively compete in a bid while keeping the profit in the subsequent phases.

The application of these techniques in real-world problems within this industry presents numerous challenges, mainly in developing a functional RL system that can be effectively implemented and utilized by these oil and gas companies. This work is a proof of concept to demonstrate the feasibility of using the RL approach for decision-making in this industry. In this sense, our findings highlights how helpful and valuable these set of tools can be in decision-making. From a practical perspective, by the end of this work, we aim to demonstrate how strategic acquisition of information assets in advance can benefit firms and expose the limitations of the rigid, inflexible, and traditional investment patterns in information assets prevailing in the current practices and structure of the industry.

Beyond this introduction, the next section presents (i) the literature review, followed by section (ii) research goals and investigation; Subsequently, we discuss section (iii) the methodology; the results are presented in section (iv) and (v) we finish with conclusions, references, and appendix.

Figure 1: Example of information Asset used in O&GI



Subsurface Structure Visualization: Mexilhão Field Seismic Data with AI-Generated Horizon Interpretation (SeisBAi Software)

CHAPTER 1

Literature Review

In this work, we deal with joining three distinct research fields. The first is the area of investigation of corporate investments. Within this large area, we are analyzing the decision to invest corporately in the oil industry. Segmenting these investments into categories such as upstream, downstream, and within the upstream in the production and exploration sectors is possible. Our interest is on the latter level, exploration investments. However, the literature generally addresses this topic from a more macro view. We will use these references as a basis in combination with specific exploration work to create a theoretical framework that is as robust as possible for this topic.

The second topic we cover is the practical analysis of asset economic valuation. By combining this with the first field, we can generate investment strategies that are theoretical and applicable in real-world scenarios. This approach allows us to objectively present the expected value that each asset should add within a strategically designed investment portfolio.

The third and final theme we explore is the potent application of ML techniques to optimize decisions. In this context, we recognize that architectures with DRL are the most promising for overcoming the challenges posed by complexity in terms of scenario, economic possibilities, and the number of interacting agents. These techniques serve as an extension of the mathematical structure of game theory in the context of economic decisions, holding significant potential for the oil industry.

The following lines present some initial references as a starting point for elaborating the literature review. We understand that this topic aims to position our research within existing areas, making it easier to identify what has already been developed, what is ongoing, and what can be added. Hence, we can observe our contribution to related fields more efficiently and leverage previous researchers' efforts.

1.1. Information Assets and Valuation

Before designing experiments with RL in investment decision problems, it is essential to establish the concept of information assets and bring some ideas about the valuation for this kind of goods. With the end objective in mind, designing experiments with RL in investment decision problems, it is vital to establish the concept of information assets and bring some ideas about the valuation for this kind of goods used ordinarily.

In this sense, two canonical references are the books named *Economics of Information Technology: An Introduction* [**5**] and *Information Rules: A Strategic Guide to the Network Economy*[**6**]. Both discuss the role of information goods in the new digital economy era that was

growing fast at the end of century XX. The main topics covered in the first one include product customization, pricing, versioning, bundling, switching costs, lock-in, economies of scale, network effects, standards, and systems effects, all themes within economists in the economy of information. On the other hand, although the second one covers topics similar to the first, it focuses on a new set of principles to guide business strategy and public policy in this "New Economic World" of digital products. We can benefit from both in our initial journey to define the concepts of information assets. From an economic perspective, information goods can be digitized (encoded as a stream of bits) as information [**6**]. Still following this author, the value of information can differ for each consumer. Certain information holds entertainment value, while other types possess business value. Regardless of the specific source of its worth, people are often willing to pay for information. Numerous strategies are available for information providers, as consumers vary widely in how they perceive and value different types of information goods. While consumers are willing to pay for information, producing and compiling it can be expensive. The cost structure of an information provider is somewhat atypical. Since this unusual cost structure drives competition in information markets, this implies an information strategy ensemble for consumers and suppliers.

It is worth highlighting some idiosyncrasies of information goods. The article [**7**] named *Measuring The Value of Information: An Asset Valuation Approach* is a fundamental reference to understand the concepts of information assets and prepare the terrain to relate this topic with investment strategies. Exploring this paper, we found a list with seven key properties of this kind of good that we will present next.

These key properties are: (i) *information is (infinitely) shareable*: this means that the information can be used for an undetermined number of users without consuming the good. One contrast noticeable is that the sharing of information tends to multiply its value, but replication does not increase its value because no "new" information is created; (ii) *The value of information increases with use*: most resources exhibit decreasing returns to use, which the depreciation of the goods can represent. However, information increases in value the more it is used, yielding increasing returns to use; (iii) *information is perishable*: the value of information tends to depreciate over time. The speed at which it loses value depends on the type of information. The information has a cycle life, where the value varies in time. For geologic or geophysical databases, this is a reality. One data set can change the value suddenly in the function of new auctions bid in some geological area or if there is a confirmation that an area does not have resources.; (iv) *The value of information increases with accuracy*: In general, the more accurate information is, the more practical and, therefore, valuable it is. Inaccurate information can be very costly to an organization in terms of both operational errors and incorrect decision-making. The accuracy required depends on the type of information and how it is used. The quality-cost-availability compromise is our core issue; (v) *The value of information increases when combined with other information*: Information generally becomes more valuable when compared and combined with other information. For example, customer information and sales information are each valuable information sources. However, linking these two sets

of information offers significantly greater value from a business perspective. In most organizations, the lack of integration of information in operational systems is a significant impediment to producing decision-support information; (vi) *More is not necessarily better*: in most cases, the more of a particular resource you have (e.g., finances), the better off you are. A standard management problem is deciding how to allocate limited resources among many competing causes. However, with the increasing use of information technology, information is anything but scarce. The biggest problem in most organizations today is not the lack of information but the overabundance of information. In practice, decision-making performance decreases once the amount of information exceeds a certain optimal point; (vii) *information is not depletable*: most resources are depletable, which means the more you use, the less you have. However, information is self-generating, which means the more it is used, the more information is created. This is because new or derived information is often made by summarizing, analyzing, or combining different information sources. The original information remains, and the derived information is added to the existing asset base. Fundamentally, therefore, information is not a scarce resource. Techniques like data mining are explicitly used to generate new information from existing data holdings.

To discuss point (ii) deeply, we can read [**8**], which evaluates profits seller in both contexts with the share of information good and where the good is used exclusively by one consumer. The authors' conclusions are unexpected and point to the dependence on the distribution costs from the consumer's side with the cost to produce an additional unit. The profit increases for the first case and decreases in the second one. A forward step is to analyze strategies for pricing information goods. One example of this can be found at [**9**], which studies the bundling strategies for information goods considering a level of decreasing value in agreement with consumer preference. The main idea is that the value of information goods declines with the amount consumed, which puts the importance of finding the optimal bundling marketing strategies in place. The research concludes that if the level of decreasing value is high, the bundling strategy is approximately optimal. While the value decreases quickly, bundling tends to be sub-optimal. These two papers are references that bring us to the complexity of the information assets economy and aim to show how the strategic approach for consumers and firms can be a positive differential in terms of utility maximization.

Advancing the theme, we can now distinguish the information good of an information asset based on the use or purpose of use. To do that, we can recur to [**7**]. To formally define information as an asset, three characteristics need to be present: it has service potential or future economic benefits, is controlled by the organization, and comes from past transactions. Firstly, the benefits may stem from utilizing or selling the assets. Information fulfills this role by enabling the delivery of services and supporting effective decision-making. Secondly, "control" refers to the organization's ability to benefit from the asset while restricting or regulating others' access to those benefits. Information also meets this criterion; if an organization possesses information, it has exclusive access unless it sells or shares it with another party. Finally, this means that control over the asset has already been obtained due to past transactions such as purchases,

internal development, or discovery. Information also satisfies this requirement. Information is usually collected as the by-product of transactions that have occurred (internal development) or maybe the result of a purchase (e.g., a proprietary mailing database) or discovery (e.g., through analysis of data).

With the concepts about information assets placed, we drive our attention to valuation aspects of the information regarding the particular topic of oil exploration and their particularities. To start this evaluation, a technical reference [**10**] named *Geological exploration theory for large oil and gas provinces and its significance* to have a base to understand geological concepts in the background of the thesis. In the oil industry, the value generated by exploration research is a set of digital information representing a geological model with probabilities of the presence of oil in the mapped reservoir in the model and the estimation of the volume recoverable. The Wildcat Well aims to adjust these models by refining the interpretation and permitting the continuous evaluations of expected value in the compass with the entrance of new information. On the technical valuation side, we will get the basic knowledge from the classical book *Investment valuation: Tools and techniques for determining the value of any asset* [**11**] by the author Damodaran, an unquestionable reference in this field.

The value of a lead[1] is directly linked with the quality of data used to create the model and the ability of the interpreter to give a geological meaning to this data. The work of [**12**] is helpful to connect the value of quality of Information in Exploration based on the reliability of 3D seismic data. The exploration deals with high uncertainty mainly due to the lack of information about the target lead. The seismic data is a very costly information source, and its value should be justified before applying these expenditures. We can find a method for obtaining the value of a 3D seismic, giving some parameters and phases (decision tree) for a Net Present Value (NPV) analysis and its possible leverage from there as a baseline to our RL models.

The real options approach is the right direction for the value analysis. To dive into this field, we find a full explanation in [**13**] how the decisions process in the O&G Exploration could be profitable using the flexibility and options of waiting. In the same topic, we also found the work of [**14**], an overview of natural options to evaluate exploration and production assets.

Complementing the valuation branch, two more general explanations of the oil and gas sector. Discussing the uncertainty and risk analysis as [**2**] does in oil exploration and production is essential. Still, the article [**15**] treats some methods to support decisions in the upstream segment. The reading of these works has helped in a broad sense while giving the macro view of the industry from the point of view of the valuation aspect.

To finish this topic, we take in advance the base idea of reinforcement learning techniques, where the reward and return need to be designed. The classic NPV and *Real Option* methods are indirectly applied to define these parameters of RL once the Time Difference methodology in the background of RL incorporates in some way these concepts. Beyond that, the environment and action spaces will also be determined based on the ideas presented in this section. In the

[1]Lead is a spatial region in a sedimentary basin with the potential to have oil or gas.

next section, we will cover the corporate investments in the exploration segment to complement all parameters of the RL experiments.

1.2. Investments in O&G Exploration

In the oil and gas industries, various types of data are collected from the surface and subsurface to understand the hydrocarbon potential. This is a crucial aspect. This industry is intensive in capital, mainly in equipment that interacts with the environmental touching, cracking, drilling, collecting, and modifying the nature[2]. There is an extra issue for exploration phases compared with other stages: the effort in capital and environmental impact could be in vain due to the intrinsic high-risk level. For instance, a seismic survey or a wildcat well can result in no oil discovery. With the emergence of IT, which affects the industrial sector across the board, the division of investment between physical and digital capital has shifted from the former to the latter. The perception is that digital capital is ultimately a substitute for physical capital since an efficient allocation based on digital assets can reduce the need for investment in physical goods. An example of this concept can be seen in recent oil field development projects, where investment in geological information is intensified to reduce the number of wells and platforms.

Considering aspects such as safety, environmental, and social, it can be said that an industrial strategy based on a higher soft/hard investment ratio will have a better result both in market competition and with the Environmental, Social, and Governance (ESG) challenges posed to this industry. However, no matter how efficient the oil industry is, it still has a minimal share of 'hard' investment. On the other hand, soft investment can be very costly and make a project nonprofit. There may be an optimum decision point for allocating resources to one type of asset over another. In this study, we propose using ML to create decision-support models for investment strategy within the exploratory process. The initial approach is generic and can be applied to other industrial segments, but we will emphasize the oil exploration model.

The O&GI is part of the energy sector, presenting a share in the energy mix that varies between 53% today and between 50–58% in projections for 2040 [**16**]. Therefore, oil and natural gas will be essential sources in meeting the world's increasing energy demand in the future. The future energy supply share between gas, oil, electricity, coal, and renewable energy can change [**17**] regarding many scenarios with a direct impact on the investment strategies of O&G companies. The major companies in this offshore industry are responsible for most of their investments in exploration and production, where the top 10 corporate investors represent up to 1/6 of total world investment, as shown in [**4**].

The research about the effects of the oil price shocks under the stock market returns of O&G companies has a broad literature [**18**]. Still, the studies of the determinants of investments seem to be less explored as appointed by [**19**] that mentioned the scarcity literature studying the uncertainty of oil prices on firm-level investment. Previous analyses in the O&G corporate investments considered distinct approaches and parameters. The research generally investigates

[2]For ecological sense, the investments can be hard (direct impact as drilling) or soft(indirect impact as a seismic survey), where the second it is not clean. However, it is less dangerous or damaging to society and the environment.

the influence of oil price changes on corporate investment, the impact of oil price uncertainty on investments as [**20**], or the economic policy uncertainty on corporate investment [**21**].

On the other hand, we find studies on investment decisions by oil and gas companies, including field-specific variables as the size of the oil and gas reserves beyond the price of oil and the oil price volatility as [**22**] or as we can see in [**23**] investigation applies advanced methods, e.g., system dynamics model to analyze and forecast the upstream investment scale and structure for oil company [3]. The variables include yield, production capacity, and oil and gas reserves, as well as expected oil and gas yield, expected production capacity, planned increased production capacity, planned production capacity at the end of the year, and planned increased recoverable reserve. The investigation studies the relationships between the annual increased proven reserve and annual investment in exploration wells, accumulated increased proven reserve, and accumulated investment in exploration wells. The forecast research assumes some scenarios with increases or decreases of some variables to compose the analysis.

The paper [**24**] investigates investment scale and structure decisions in the upstream sector using system dynamics theories to evaluate the actual results from an oil company's practice in China. Other articles, such as the [**25**], discuss a more general view of investment decision-making in the upstream oil industry. In contrast, [**26**] analyzes the oil price impact on the investments and production of U.S. industry, and [**27**] treats the historical energy consumption and future demand.

1.3. Strategy Analysis and Machine Learning

The study of RL presents a vast literature with new material constantly emerging. Given this multitude of possible sources for reference, we will stick to the most solid and necessary content to carry out the work planned in the thesis. So we can start with the main book written by Sutton & Barto entitled: "Reinforcement Learning: An Introduction" which can be found in [**28**]. This is a secular book for everyone who enters the RL world. Other materials will complement our list, but for a while, this book will be sufficient to start. In the opposite space, we have an article that is an inspiration and a solid basis for our work named "Revealing Robust Oil and Gas Company Macro-Strategies using Deep Multi-Agent Reinforcement Learning" by [**29**]. This paper presents high-quality results, including the precise DRL architecture description and figure and graph designs. We will use these two references a lot throughout the entire project. To complement this work, we can read [**30**], where the same topic is treated with a slightly different approach.

The use of ML in applications in the energy industry has become increasingly large and diverse in its segments like in [**31**] who work on practical applications of ML methods in the energy areas, including the petroleum supply chain, steel-making, electric power system, and wind power. In another example, in[**32**], we can observe this methodology being used to find predictors of the price of oil. This type of application is frequently found in financial analysis to forecast stock prices and returns. In economics, specifically in competition analysis, there

[3]This paper presents detailed components of total investments in some diagrams, that can be consulted in case of understand deep the inner distribution of investments in upstream link

are several works in progress, mainly considering the search for Nash equilibrium in systems modeled by Bertrand and Cournout games, among others. A practical example, the article [**33**], takes an RL model to assess firms' market power under auction-based energy pricing. Although there are already several works in the bid auction research, there are green fields for ML applications. In some way, our research will be close to this type of study. Still on the tenders [**34**] developed studies to aim bidders' recommendations for public procurement auctions using ML.

To stick closer to our theme, the work [**35**] brings a lot of valuable methodologies to deal with economic problems using ML into the portfolio strategies. The primary author's idea is to use DRL power to improve portfolio management, which correlates with testing investment strategies. Diving more in this field, we found an updated reference in [**36**], with the work on optimization portfolio with machine learning techniques.

Additionally, the workers as the [**37**] a review of the combination of two main fields, in this case, energy and finance. Interestingly, these advanced techniques have brought new perspectives to many distinguished areas. However, beyond that, they have also been able to integrate fields previously analyzed together. Even so, we chose this methodology because we understood that complex business strategies combine different facets of the same problem to make the optimal investment choice. These references corroborate the decision, helping to know how to use the tools to ensure an excellent multivariate analysis.

Finally, we can find on [**38**] an indication to use DRL for decision-making regarding an optimization based on socially responsible investments to select assets to include in a portfolio. This work reflects the pressing need for companies, especially oil companies, to root the dimensions of environment, social responsibility, and governance in the decision-making process and portfolio formation.

CHAPTER 2

Research Goals and Investigation

In recent years, the oil industry has changed rapidly in terms of processes and routines to adapt to changes in society's behavior, but above all, to incorporate significant technological advances that have allowed greater efficiency and returns for companies in this sector industry. In this context, in a competitive market, digital transformation and the direction of processes towards a data-driven vision are in full swing for some oil companies through strategic internal channels for upgrade efficiency in exploration and production. Internal decision forums can be established to meet the needs of these programs that depend on input data set in the $E\&P$ chain, which aims to deliver the optimal geophysical solution for executing the Strategic Plan. In this plan, one of the main bottlenecks is analyzing the economic value of the various possible solutions. This process helps decision-making obtain geophysical information, considering quality parameters, availability time/life cycle, costs, regulation, and environmental issues. Some methodology creation initiatives are underway, among which the incorporation of artificial intelligence techniques stands out, as well as the methodology with utility functions of multi-factorial information, whose development at the exploration data acquisition department is innovative for the oil industry.

Our initial research proposal considers two possibilities: i) incorporation of advanced economic valuation techniques, such as multi-factor utility functions and real options theory; ii) use of Machine Learning and Deep Learning to extract regularities in the vast number of parameters that relate the links in the $E\&P$ chain, such as the relationship between the countless acquisition and processing parameters and the difference between the predicted result in the "digital rock" versus the result obtained from drilling the physical rock, to measure the impact of the first set on the second. In addition to signaling whether a given geophysical solution is viable from the perspective of value, this investigation encourages the creation of new databases, either through the integration of information between different areas of exploration or through the generation of information from the existing data or even for allowing the redesign of processes to a data-driven-friendly standard.

2.1. Thesis Objectives

This project addresses a critical topic for oil companies: the quantitative analysis of parameters traditionally evaluated qualitatively in Exploration. For example, what is the real impact of a quality delta in geophysical information on the exploratory risk of a given area? What about the results of the delimiting or extension wells? And in anticipation of production? The qualitative/subjective analysis results from the complexity of answering this question. However, with

the recent evolution of computational techniques, we understand that moving forward and gradually switching oil companies' processes from subjective traits to more objective ones based on historical data and statistical models is possible. Furthermore, in the valuation of projects at oil companies, we observed the majority presence of traditional cash flow techniques and little development in using ML or other more robust methodologies that bring flexibility in the decision-making process, a critical factor in the super dynamic environment of the petroleum industry. In summary, our general objectives are described as follows.

2.1.1. General Objectives

The general objectives are i) find out the strategy for corporate investment in information assets in the exploration of the oil and gas industry; ii) Establish a standard decision analysis process based on the evaluation of different types of exploration inputs; iii) understand how the information volume and quality impact the chances of success of an exploration area; iv) Investigate how each kind of information can be complementary or substituted in an exploration chain in a strategic sense; and finally v) Analysis of the substitution of information by incorporating measures of ecological risks to find assets that rely on environmentally sustainable strategy.

2.1.2. Specific Objectives

The specific objectives are i) Develop economic valuation methodologies using Artificial Intelligence resources, beyond the make possible uses of Multi-factor Utility, and extend to valuation methodologies based on Real Options and ii) As a by-product, we have the basis for automated quality analysis of seismic information through Machine Learning.

2.2. Research Hypotheses

Currently, the main strategy for investing in information assets in exploration considers the investment versus risk relationship in a "ladder pattern" over time, where it is expected to obtain a greater knowledge of exploratory areas to increase investment in higher quality information. The rationale behind this strategy is that exploration is an activity with high intrinsic risk and that greater early investment can turn into economic losses if production development is not pursued.

However, an alternative to this strategy is to anticipate investing in assets of higher quality, even with some risk. The justification for this alternative is that, if successful, the entire exploration and production development chain benefits from more assertive geological models. In addition, there is an environmental benefit, as it prevents ships from returning to collect new data, reducing the fold of surveys in an area. Based on these arguments, we present the central research hypothesis:

H1: It is possible to obtain an alternative strategy for investing in information assets that presents a higher return considering different economic scenarios and broader aspects such as environmental impact.

H2: The second hypothesis derives from the first and asserts that the alternative strategy is robust, which means it systematically presents better results.

2.3. Research Contribution

At the end of the research, once the objectives defined here have been achieved, it is expected that technical and managerial decision-making regarding which technology and when each piece of information should enter the exploration process chain will be more assertive, resulting in a reduction in exploratory risks, costs, environmental impacts and value maximization for the interested oil company. Additionally, modeling the decision-making process using ML is a recent and rapidly developing research field in economics for which applications in different contexts can generate improvements and insights for other researchers.

CHAPTER 3

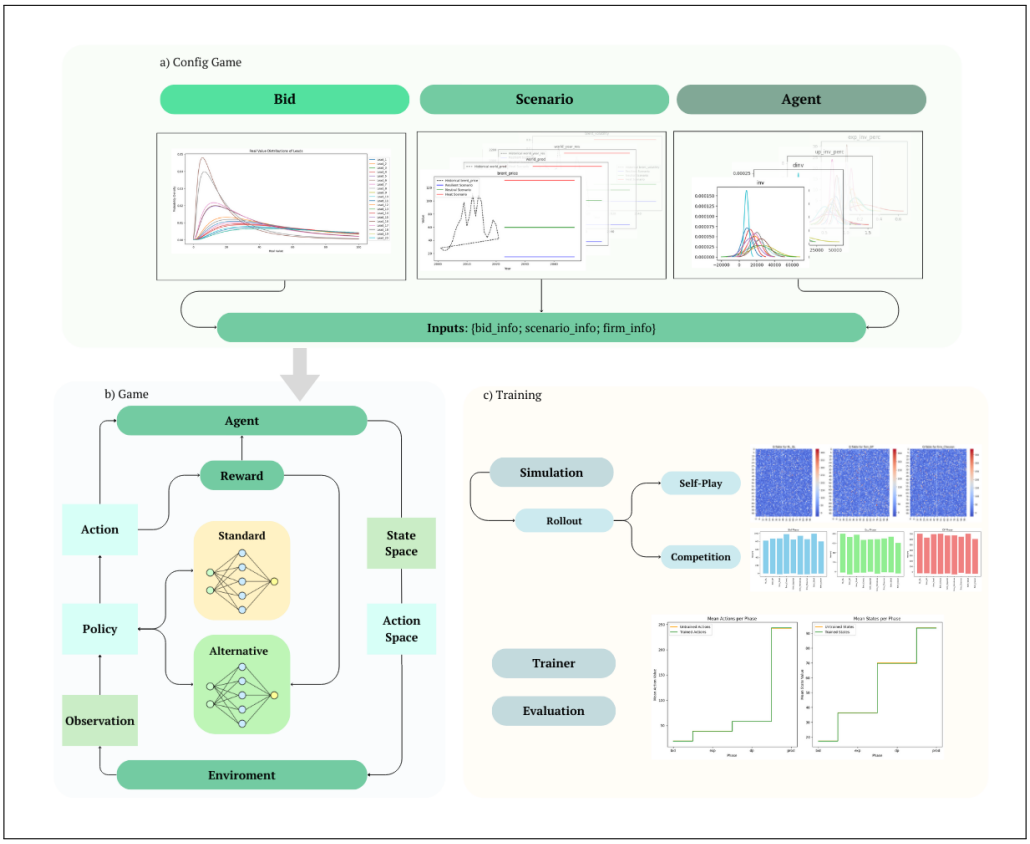
Methodology

In this topic, we briefly show the methodology used in this work. To do this, we present some basic elements of RL and the code structure necessary to carry out the experiments. In this way, the first subsection provides an introductory overview of RL, and the following section provides information on code tools and utilities that we used to execute the project. Our basis to define the methodology is found in [**29**], where authors modeled a multiple-player war game with the agents being the majors of the O&G industry. The environment considers transition energy constraints in possible future scenarios, and the reward is the profit at the end of the experiment. To obtain the optimal strategy, they use a DRL. In this paper, we observe an architecture with three principal components. Begin with the action space activated based on endogenous and exogenous economic variables, such as production, cash flows, trading, low-carbon asset auction bidding, and capital allocation. The game advances to the next stage after the allocation has been taken from the action space. Upon reaching the end game, a new energy scenario is chosen to alter the dynamics of the next game. In second place, oil companies are trained in deep reinforcement learning. The learning companies collect game experiences in the war game via game state observations and the reward signal to update their strategy. Frozen companies are chosen by the multi-agent learning mechanism and compete against the learning company. In the last stage, they use a league training system involving two learning companies to address game-theoretic challenges. Leading companies do not have strategy constraints but vary in initial asset distributions and are initially trained against copies of themselves. Exploiters are added as opponents in subsequent iterations. Exploiter companies are constrained to specific strategies. This work is one of the few references in the field that have a complete architecture for oil and gas analysis, unhappiness the code is not shared.

In our research, the rationale is quite similar; what changes is that we will evaluate investment strategies in information assets in the upstream phases. The agents are still oil companies, and the input information is the previous investments and internal and external variables of these firms. The environment will first be a model of fundamental conditions in four phases: bidding, exploration, development, and production. The actions can be to invest in higher- and lower-quality information assets, making this investment one of the determined phases. The environmental conditions for these decisions must include exploratory risk and success rate information throughout the industry chain. To correctly design the game, it is necessary to elucidate the main concepts of RL as shown in Figure 2.

Our architecture consists too of three parts. The first is to create the inputs for bid and scenario generating the lead competition and market conditions, which simulate futures and the agents configured from real data throughput the profiles of firms. The second part is named

Figure 2: DRL Architecture for Strategies in the O&GI



Note: a) The game consists of three key components: Bid that generates market conditions, Scenario which simulates futures, and Agents that are configured from real data; b) Game dynamics include Action, Policy, Environment, State, and Reward, where agents interact with the environment, receiving rewards, and their performance is compared to an untrained agent; c) The training process refines the agent's policy through simulations, with experiences updating the agent's strategy for better performance across scenarios.

the game dynamics. It includes classes such as action, policy, environment, state, and reward, where agents interact with the environment, receive rewards, and perform better than untrained agents. Finally, we have the training process that refines the policy through many simulations, with experiences updating the agent's strategy for better performance across market conditions and leading competitions.

In the next section, we bring a brush to the topic and explain how to format the code for this type of technique in a simplified view.

3.1. Reinforcement Learning

Reinforcement Learning is a machine-learning technique based on try-and-error philosophy where an agent makes decisions to select an action in an environment and receives a reward signal each time that the new state is better than the old state. The RL algorithms have essential concepts in former backgrounds like (i) agent, (ii) actions, (iii) environment, (iv) state, and (v) reward. We will discuss which one of them forward. Beyond these structural concepts, some other components and parameters emerge to include the particular aspects of each problem

forming the game. The inputs for RL are the parameters of the concepts above. The output can be an optimal policy maximizing rewards along the pathway or just the return optimum. The return is defined as the sum of all rewards weighted by discount rate γ, with a similar mean of the economic discount rate.

There are two types of reinforcement learning algorithms: model-based, in which an optimal policy is explicitly determined on the basis of a learned model of the environment, and model-free, in which an optimal policy is derived entirely from trial-and-error experiences. From another point of view, the RL can be value-based, where agents learn the value or action-value pair function to generate the optimal policy generated implicitly, or policy-based, which states the policy directly using a parameterized function. An additional policy classification is the Actor-Critic RL, which mixes the value function with the policy-based statement. RL can also be classified into on-policy and off-policy methods. In an on-policy approach, often described as "learning on the job," the algorithm learns about the policy π by directly interacting with and sampling from the policy Π. Conversely, in an off-policy approach, like "looking over someone's shoulder," the algorithm learns about the policy π by utilizing experience sampled from a different policy μ. Notable examples of off-policy methods include *Q-Learning*, *Deep Q Network* (DQN), and *Deep Deterministic Policy Gradient* (DDPG). On the other side, we consider on-policy methods that apply techniques such as *Asynchronous Advantage Actor-Critic* (A3C) and *Proximal Policy Optimization* (PPO). The other thing to mention is that the RL's action and state spaces can be discrete or continuous, further diversifying the classification of RL problems. For a more comprehensive understanding of these concepts and the broader taxonomy of RL can be found in [**28**].

If seen from a coding oriented perspective; since the state and action space is discrete, we can iteratively estimate the action value function using Bellman equation as:

$$Q_{i+1}(s,a) = R + \gamma \max_{a'} Q_i(s',a') \tag{3.1}$$

This helps us to determine the new value that should be assigned to that action from each state by looking at both the expected reward and its estimated future rewards. Persistence and dedication to this field are apparent through their iterative process. As $i \to \infty$, it converges to the optimal action value. All this means is that the agent must explore the state-action space in a systematic way, updating its estimate of $Q(s,a)$ at every step until it comes in alignment with the optimal action-value function: $Q(s,a)$. However, it becomes impractical to comprehensively explore the entire state-action space when the state space is continuous. As a result, it is equally challenging to accurately estimate $Q(s,a)$ until it converges to $Q^*(s,a)$.

Deep Q-Learning: One of the critical parts in Deep Q-Learning is to estimate action value function, $Q(s,a) \approx Q^*(s,a)$ and this problem statement is handled using neural network that we call as Q-Network. We use the Bellman equation to govern how the weights of this network are updated iteratively in order to minimize mean-squared error. However, in reinforcement learning using neural networks tends to result in highly unstable methods. There are two main

techniques used to address these issues that are explained below; Target Network and Experience Replay. In every iteration of training, we train the Q-Network's weights so that error between the predicted and target values is minimized. Here the Target value will be:

$$y = R + \gamma \max_{a'} Q(s', a'; w) \tag{3.2}$$

where w represents the weights of the Q-Network. Consequently, the goal is to minimize the following error at each iteration:

$$R + \gamma \max_{a'} Q(s', a'; w) - Q(s, a; w) \tag{3.3}$$

The Q-Network target values are indeed defined in terms of the original Q-Network, and computing those without modification would involve updating the targets down the same weights lots of times for a given starting state. This causes oscillations and instability in the training procedure. CNNs fail to converge because the target,y, and care are always moving. The target Q-Network is just another neural network with the same architecture as the original Q-Network — we use this network to alleviate this issue. The target $\hat{Q}$-Network is updated less often and remains a slower-moving target. Then, the error we seek to minimize is now in terms of this target $\hat{Q}$-Network:

$$R + \gamma maxa`hatQ(s`, a`; w^{-}) \tag{3.4}$$

, where w^{-} are the weights of target $\hat{Q}$-Network and w are the weights from original Q-Network. This offsets the training so that the volatility in target values is reduced.

where w^{-} represents the weights of the target $\hat{Q}$-Network, and w represents the weights of the original Q-Network. This technique works to stabilize the training process by decreasing fluctuations in the target values, making updates to the model more uniform over time.

We need some tools and infrastructure to run the code.— Python interpreter (v3.9 or higher) The language we use to write the code is Python, but shortly may have to resort to some others. By using a Python-compatible framework, codes can be written in either notebook format or the. py extension. A framework compatible with Python is used to write codes, which can be in notebook format or the .py extension. The systems already in use are *Visual Studio Code* (*VS Code*) and *PyCharm*. Several open-source libraries are used for both data processing and visualization. Among the main libraries, we can anticipate some fundamental as *Numpy*, *Pandas*, *TensorFlow*, *PyTorch*, *Gym*, and *Matplotlib*. The Gitlab application is used to store and share the codes.

We are currently working on development of the algorithm which runs along with this empirical analysis. This is done to align both aspects in the subsequent phase of this research. The purpose of this pseudo-code presentation is to delineate what must be demonstrated in transitioning the game from a model inspired to model driven.This version will allow for the execution of experiments in an optimized manner. The primary references that aim to develop architecture and codes are [**29**], [**30**], and [**39**]. For Deep Reinforcement Learning issues, we address the authors [**40**] and [**41**]. Further details about this model will be provided in the following section.

ALGORITIMO 3.1. *Deep Q-Learning Algorithm*

DQN Pseudo-Code

Require: $n \geq 0 \vee x \neq 0$
  $D \leftarrow np.zeros(N)$
  $w \leftarrow np.random(N)$
  $\hat{Q}$-*Network with* $w^{-} = w$
Ensure: $y = x^n$
  for $episode\ i$ *in* M ***do***
    $S \leftarrow S_i$
    for $t = 1$ *to* T ***do***
      $A_t \leftarrow A(S_t, \epsilon)$
      $E_t \leftarrow E(A_t)$
      $R \leftarrow R(S_t, A_t)$
      $S_t \leftarrow S_{t+1}$
    end for
  end for
  while $N \neq 0$ ***do***
    if N *is even* ***then***
      $X \leftarrow X \times X$
      $N \leftarrow N/2$
    else
      $y \leftarrow y \times X$
      $N \leftarrow N - 1$
    end if
  end while

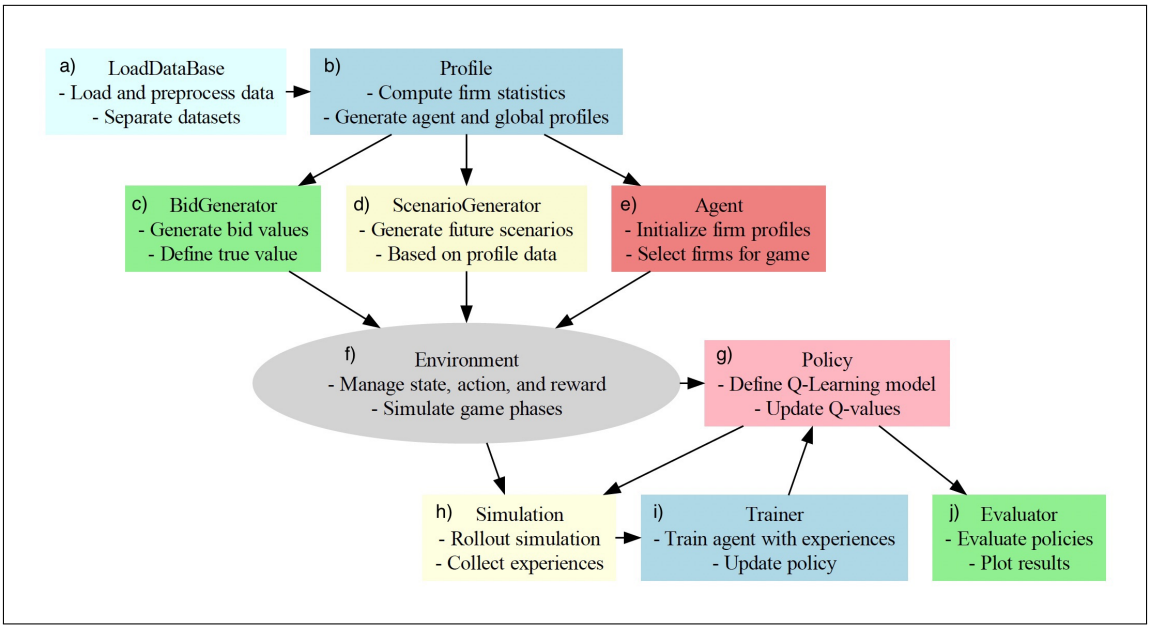
3.2. Model

Upon introducing the fundamental concepts of RL, we can now draw a correlation with the issue of information assets in the marketplace. To succeed with the base model, it is imperative to define a comprehensive game, with all parameters meticulously designed to represent the hypotheses under investigation. The initiation of all games involves the selection of the player mode and skin. Let us proceed with that, looking for the Figure 3. This figure presents the general flow with a representation of each class in the RL code. Before detailing the classes, we establish our model's main components and respective economic attributes.

3.2.1. Agents

In this game context, several potential agents can be identified. The primary agents are the Oil and Gas (O&G) firms. However, other entities such as regulatory bodies, environmental or government agencies, social representatives, suppliers, and academic researchers also play significant roles. For the time being, our focus will be solely on the O&G firms. These firms were selected based on their ranking in offshore investments over the past two decades. The results of this selection can be found in Table , referenced in section 4. According to the IHS database for seismic projects, there are more than 450 companies. However, the top ten represent up to 40% of total investments in some years. Gathering data for such prominent

Figure 3: RL Flow for Investment Strategies in Information Assets



Note: a) Load data and separate firm-level and global datasets. b) Computes each firm's statistics, generating agent profiles and global scenarios. c) Generates bid characteristics and true values for the game simulation. d) Generates future scenarios for the simulation based on profile data. e) Initializes profiles and selects firms to participate in the simulation game. f) Manages each game phase's states, actions, and rewards. g) Implements the Q-Learning algorithm and updates Q-values based on simulation results. h) Simulates the game and collects experiences for training. i) Trains the agent using collected experiences and updates the policy. j) Evaluate both trained and untrained policies and plot comparative results.

companies is a challenging task. Therefore, an attempt has been made to balance the number of firms included in the study and the representation of the entire market.

These top-ten companies are the most significant in offshore investments for the upstream chain, without loss of generality. Future analyses could include more firms, thereby expanding the database. Formally, we call the set of firms as:

$$\Im = \{f_1, f_3, f_3, ..., f_9, f_{10}\} \tag{3.5}$$

3.2.2. States

Geometrically, the state is a vector of outcomes that exist only from an environment, i.e., E. Consider the environment similar to a game background, for example. In this regard, the state will resemble a screenshot of one moment during a play-through. This screenshot, or state, provides a snapshot of all the relevant parameters at that particular time. In this way the set $S = \Im \times E$ of states is given by:

$$S = \{firm_i, [phase_k, \, inf\_asset_{i,q}]\}, i \in range(dim(\Im)) \tag{3.6}$$

3.2.3. Actions

The action space is defined by two variables: the magnitude of investment in information assets, which depends on quality, and the phase in which to invest. We define action space as follows:

$$A = \{phase_k, inf\_asset_{i,q}\}, \quad for\, k \in [1,2,3,4], i \in range(dim(\Im)) \quad and \quad q \in Q. \quad (3.7)$$

Both could be discrete or continuous. Nonetheless, we start with a discrete-time model for the dividend phase characterized by constant time soot and switch to a continuous representation in investment in information assets.

Investment in information assets influences the risk associated with a lead. Each year, the highest and lowest technology quality parameters determine the boundaries for investment in information assets. The cost of information is directly proportional to its quality and uncertainty.

3.2.4. Reward, Discount Rate (γ) and Return

The phase changes can increment the firm's value, reduce the business risk, or both. The rewards W are different for each phase. In the initial model, $k = 4$ is used for phases named bid, exploratory, development, and production. The initial phase reward, denoted as W_{bid}, represents the authorization to investigate the lead area. The assets being integrated to an exploratory portfolio of values ($r_{i,exp}$, inv_exp_i) are undergoing revaluation through this process. We consider the subsequent phase at which point W_{exp} denotes a reward entitlement when there is an affirmation that this lead can be profitably to serve according the values ($r_{i,dp}$, inv_up_i). This W_{dp} works as the factor that strongly promotes investment in physical assets and at the same time mitigates some of risks ($r_{i,prod}$, inc_res_i) in 3 rd phase. The last term W_{prod} refers to the extractable resources financial worth and their contributions raising the value of firms, $share_price_i$, $revenue_i$ or$market_value_i$.

The return is the value discounted by γ and $r_{i,k}$. The time frame for each k is $[1, 4, 2, 18]$ years. For these parameters, we define return R as:

$$R = W_0 + (\gamma + r_{i,1})W_{bid} + (\gamma + r_{i,2})^5 W_{exp} + (\gamma + r_{i,3})^7 W_{dp} + (\gamma + r_{i,4})^{25} W_{prod} \quad (3.8)$$

The policies are designed to optimize the exploratory portfolio during the bidding phase. As a reference, we cite [**33**] or [**34**], validate reservoirs in the exploration phase, minimize costs, and expedite production during the development phase. Ultimately, these strategies aim to enhance total dividend payouts, revenues, and profits during the production phase or to optimize the portfolio ([**35**], [**37**] and [**38**]) in the O&GC context. The incorporation of negative rewards can encourage robust strategies and realistic agent behavior. The optimal return formulation will be identified through experiments and model refinement processes. At present, we are initiating the preliminary step for the base model.

3.2.5. Environment

The environment is constituted by all the necessary variables required to configure a potential state within the game. We have the basic components as $year$ and $phase_k$, where the last can be bid, exploratory, development of production, and production. The global variables are, for

investments, tot_inv, tot_inv_up, and tot_inv_exp, representing the industry's total investment, and segmented values for upstream and exploration. For the operational aspect, the variables tot_prod, tot_res, tot_inc_res to the total production, total reserves, and the total reserve increment by year. For the market domain we consider the oil_demand, oil_price, and oil_price_vol In the firm's domain, we have:

$$
\begin{aligned}
E = \{&\text{year, phase}_k, \\
&\text{tot\_inv, tot\_inv\_up, tot\_inv\_exp, tot\_prod, tot\_res, tot\_inc\_res,} \\
&\text{oil\_demand, oil\_price, oil\_price\_vol, inv}_i\text{, inv\_up}_i\text{, inv\_exp}_i\text{,} \\
&\text{prod}_i\text{, res}_i\text{, inc\_res}_i\text{, share\_price}_i\text{, share\_price\_vol}_i\text{,} \\
&\text{revenue}_i\text{, market\_value}_i\text{, risk}_{i,k}\text{, inf\_asset}_{i,q} \\
&\}, \quad \forall f_i \in \Im
\end{aligned}
\tag{3.9}
$$

3.2.6. Policy

The Policy should implement a reinforcement learning strategy using a Q-learning algorithm. The core is the Q-table, which is initialized as a matrix with dimensions defined by the number of states and actions. The policy converts states and actions into indices through hashing, which allows for efficient indexing into the Q-table and determines the next action to be taken by the agent, balancing exploration and exploitation based on an ϵ-greedy strategy. The Q-table is updated by applying the standard Q-learning update rule. The new Q-value for a state-action pair is computed as follows:

$$
Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right] \tag{3.10}
$$

where α is the learning rate, γ is the discount factor, r is the reward, s is the current state, a is the current action, s' is the next state, and a' is the next action.

In the SARSA, that is an on-policy, meaning it updates the policy based on the current agent's actions. This implies a slight difference in the Bellman equation for updating q-values:

$$
Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma Q(s',a') - Q(s,a) \right] \tag{3.11}
$$

We replace the $\max_{a'} Q(s',a')$ in Q-Learning, by the current action a' chosen by the agent in the state s'. The SARSA tends to be more conservative for this on-policy mode, whereas Q-Learning is more aggressive in its exploration.

For DQN methodology, we replace the state-action Q-Table with a neural network that receives the states, actions, and rewards as input, delivering a vector containing the Q-values for each action in the current state. In the input layer, the network starts with a fully connected layer that accepts four inputs for each phase in the state representation. This is followed by applying batch normalization to the inputs, improving training stability and helping the model generalize better. After this, we tested ReLU and Leaky ReLU as the activation function, which helps avoid the "dead neuron" problem when using a small gradient, even for negative inputs. This improves gradient flow during training. We have layers to do a dropout after the second

fully connected layer to reduce over-fitting. Dropout randomly sets some neuron outputs to zero during training, making the model more robust. The final layer outputs a vector with the Q-values for each possible action for each phase. These Q-values are used to make decisions based on the expected reward for each action.

3.3. Code Implementation

This code provides a reinforcement learning (RL) model applied to an oil and gas setting in a simulation environment. The main goal is to model the behavior of multiple competing firms, each operating through different phases: bidding, exploration, development, and production, to evaluate how different strategies would affect their performance. The Environment contains rudimentary components related to data loading, preprocessing and scenario generation, agent behavior, and policy evaluation. The first set of implementations is encapsulated in the *LoadDataBase* class, which loads and processes the dataset. An ExcelFileManager class that reads data from a spreadsheet, parsing it into firm-level and global data. It also includes methods to summarize the data and extract specific information related to the firms and years of operation. The *Profile* class follows, which computes statistical profiles for the agents (representing firms) and the Global Environment. Since this class produces statistics for the firm-specific and global datasets, it sets the conditions under which the Simulation will run. These classes, in addition to more relevant ones, including BidGenerator and ScenarioGenerator, which create bid values and future scenarios, respectively, also participate in the Simulation. These are used to generate the inputs for the Simulation, allowing the behaviors of various firms with respect to the Environment to be variable and realistic. The *Agent* class is designed to initialize and manage the profiles of different firms, selecting a subset of them to participate in the game simulation. It comes with support functions for profiling and running the agents inside a simulation. The simulator revolves around the *Environment* class, which acts as a container for everything about the game, its state, action, and reward functions. It attempts to mimic the sequential set of decisions that firms need to make and computes the second-round results of such actions iteratively through these stages. The Environment calculates the expected rewards and handles when an agent reaches one state to another depending on what it did. Each implementation competes against one another by virtue of its Policy class, which relies on Q-Learning and SARSA algorithms to update the strategies of the agents given their experiences in the Environment. This class has methods for selecting actions, updating Q values, and exploration-exploitation tradeoffs.

The Simulation and training process is managed by the *Simulation* and *Trainer* classes. The *Simulation* class handles the roll-out of the Simulation, collecting experiences that are then used to train the agents. The *Trainer* class coordinates the training, It loops over many episodes to update the agents' policy. Also, the *Evaluator* class can perform such performance comparison between trained and untrained policies by presenting different metrics, viz., rewards, state-action distributions, etc., to measure the efficiency of policy training. The Evaluator has functions for temporals, visualization of the Q-table, rewards, and actions, and display to learn how the behavior of the agents has evolved.

This implementation showcases a detailed approach to modeling a complex multi-agent environment using reinforcement learning. It includes setting up data, generating scenarios, and performing Simulations, which are exact work followed by training and a good evaluation process. The code was written to be modular, so testing and modification of each component could be done individually. Such attention to detail carries over to the model, providing trust in its learning capabilities... an important capability for addressing some of the complexity found in oil and gas operations.

3.4. Database

The strategy analysis using RL has the advantage of benefiting from synthetic data. Based on the philosophy of RL, the main idea is to create a math structure to model agents, actions, and environments in the totality of the possible state space. In this way, the inputs for the models are rules and hyper-parameters defined in the game's design and the tests to improve performance. Notwithstanding, using real data to validate the network's trained outputs is recommended to check if there is a correspondence with reality.

Furthermore, our synthetic database will originate from a design game supplemented with test data from exploration bids, costs of information, and data of O&G companies. The primary market for collecting real data will be global, specifically focusing on Brazil for certain experiments. However, this is an open focus, and other markets may also be considered. As the need arises, we will update the database to enhance the quality of parameters, thereby improving accuracy and performance. This will, in turn, optimize the results for the trained network. The data will be gathered from public agencies' websites that regulate the O&G sector. These data are derived and expanded from a previous work database in [**4**]. These data can be sourced from the F-20 annual reports of companies listed on stock markets or the EIKON platform in a DATASTREAM module.

As mentioned, We collect data using the EIKON platform to access DATASTREAM for daily information on stock prices of $O\&G$ companies and $Brent$ and WTI oil prices from the first weekday of January 2001 to the last weekday in December 2021. With this raw data, we calculate the daily, monthly, and yearly returns of stock price firms and oil prices and their volatility(σ_{stock} and σ_{oil}).

Additionally, we calculate the volatility using the Levi exponent. The volatility has a controversial role in explaining the variability of corporate investments, which leads us to investigate a further formulation for this variable. The volatility measures act as a proxy of the risk for both the company (σ_{stock}) and the oil market(σ_{oil}).

In the discussion section, we revisit comparing the canonical volatility and the index derived from the Levi exponent. Each company segment's production, reserve, and total investment can be located in the annual reports, denoted as F-20, either on the EDGAR database or the respective company websites. We have meticulously collected and organized the data from these reports to facilitate our environmental validation. Following this concise explanation of our data-gathering efforts, we will explore this unrefined diamond in the subsequent section.

CHAPTER 4

Results and Discussions

4.1. Results

In this section, we will analyze the results of the RL system developed based on the methodology described earlier. The analysis is carried out through a series of experiments, each comprising several rounds of the O&G game, which simulates bidding and decision-making processes from the bid phase through the exploration, development, and production. We begin with an empirical analysis of the generated database followed by the results of the RL system.

First, we start by presenting preliminary results that serve as inputs for the agent's training process. These intermediate steps, derived from the bidding classes, scenarios and agent generators, are calibrated with real-world data to ensure the simulation aligns with the reality of the O&GI, providing a realistic foundation for the agent's learning processes.

The subsequent results are related to the states, actions, and rewards observed and accumulated by the agents throughout the simulations. These outputs reflect the agents' interactions within the simulated environment. They are necessary to run the experiments and to proceed with our research. Furthermore, observing the actual behavior of firms in the market reveals a good representation of their states and actions within our model. Therefore, the reward structure was extensively refined for a long period during the experimentation phase to provide the correct functional form in the game.

In the last part of this section, we present the results related to policy development. Additionally, we also provide the Q-Table Heat Maps for the Q-Learning method and network visualizations for Deep Reinforcement Learning (DRL) approaches to enhance understanding. These visualizations represent the learned policies and the value of each action in a given state, aiding in understanding the agents' overall decision-making process. We have also included graphical representations of the average rewards per phase and policy. This information serves as evidence for testing the hypotheses outlined in this thesis. These results are vital to understanding the agents' learning process and the RL system's overall performance. We will now proceed to the next sections with a detailed analysis of the results.

4.1.1. Empirical Outcomes

In this section, we provide a complete empirical description of the database. We also analyze the data used to establish the firm and market-level profiles, along with the variables inside the environment. This step is fundamental for understanding the basic components of the simulations during the agent's training.

We aim to identify an optimal set of variables encompassing physical and financial dimensions. This will aid in constructing a structural model designed to define the corporate investments of O&G companies. Our model is predicated on three dimensions: (i)Firm-specific factors: This includes variables such as daily production, oil reserves, stock price, and volatility; (ii) Investment patterns: This dimension encompasses the total annual investments, along with the distribution of investments across upstream, downstream, and exploration activities; (iii) Market conditions: This dimension incorporates variables such as oil price and its volatility, global oil consumption, global oil reserves, and global oil production. We posit that this comprehensive set of variables will facilitate the explanation and prediction of total corporate investment for O&G companies operating in the offshore segment.

For the firm's domain, we have included the type of firm, categorizing as either national oil companies (NOCs), or international oil companies (IOCs). These domains are defined across both temporal and spatial basis. Subsequently, we have defined each feature's frequency and time interval within the primary data. Certain features, such as oil and stock prices, have a daily frequency, which is the foundation for computing the return and volatility. Other features exhibiting daily periodicity are the oil and gas production per firm. Finally, the monthly data is derived from the daily dataset through a series of transformations.

Moreover, we have accessed yearly data for reserves and investments. Regarding the temporal range, while we have found databases extending back to 2000, data is limited for certain domains, particularly those related to market analysis. The primary data consists of information collected from the firm's reports. For the spatial domain, we have utilized real-world data. However, limitations in the spatial analysis arise from the constraints imposed on the features, such as the oil and gas reserves, production, and consumption.

Table 1 presents the minimum, mean, and maximum values for the total investments, production, and reserve for top-ten oil and gas offshore investors between 2001 and 2021. The values are ranked in descending order according to the mean investment. The company Shell leads with the highest mean and maximum investment values, followed by Exxon and Petrobras. Equinor, Rosneft, and ONGC have the lowest mean investment in this top-10, while CNOOC displayed the lowest minimum investment during the same period. The maximum investment value for Shell, at $\$79.9$ billion, is an outlier due to an asset acquisition process in 2016. The annual minimum, average, and maximum time series of top-ten $O\&GC$ offshore investors between 2001 and 2021 can be seen in Figure 4.

Table 1: Firm Variables, to top-ten O&GC Offshore from 2001 to 2021

	inv (B USD)			prod (MBOE/day)			res (BBOE)		
O&G	min	mean	max	min	mean	max	min	mean	max
Shell	11.8	27.0	79.9	2.9	3.7	5.3	9.1	12.3	14.3
Exxon	12.3	25.1	42.5	3.7	4.1	4.5	11.3	20.2	25.3
Petrobras	6.5	22.2	48.1	1.3	2.3	2.8	8.8	11.3	13.1
Chevron	7.4	20.9	41.9	2.5	2.7	3.1	10.6	11.4	12.1
BP	11.2	20.1	36.6	3.2	3.6	4.0	16.4	18.0	19.9
Total	7.7	17.4	34.4	2.2	2.5	3.0	10.5	11.3	12.7
Cnooc	1.5	13.1	31.1	0.3	0.9	1.6	1.8	3.5	5.7
Equinor	2.3	12.4	23.8	1.0	1.8	2.1	4.3	5.2	6.2
Rosneft	0.4	9.0	16.9	0.3	3.2	5.8	1.4	24.4	44.9
Ongc	3.2	8.3	12.1	0.6	0.7	0.9	4.5	7.7	10.5

Note: The values are approximated to one decimal place. We abbreviate the name of ExxonMobil, and Equinor represents Statoil.

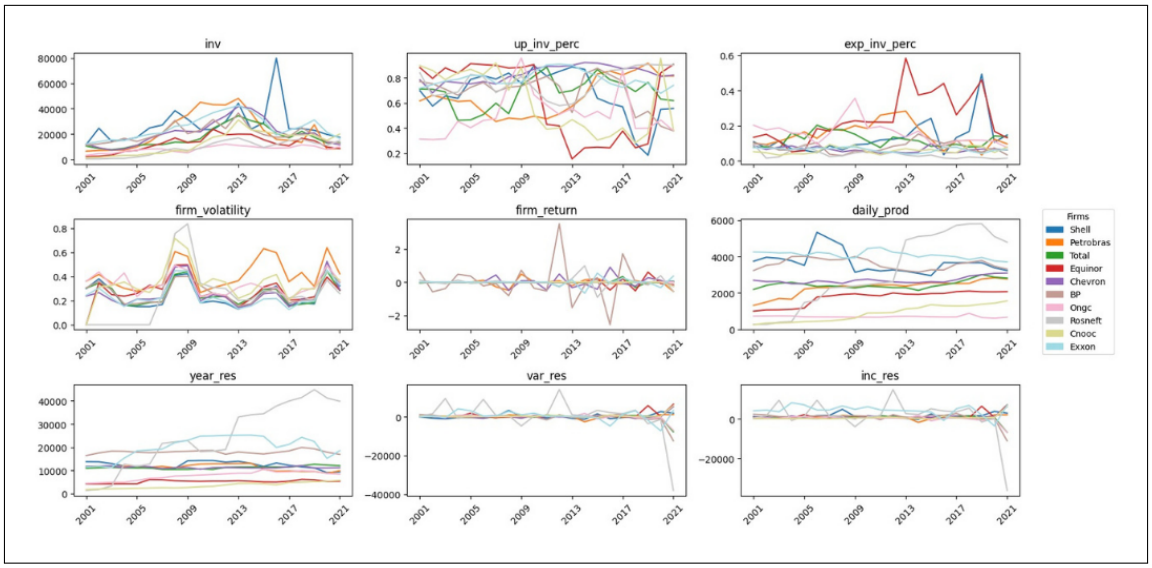
Figure 4 illustrates various firms' annual investments, production, and reserves. This graph highlights the concurrent evolution of Rosneft's production and reserves. In 2001, Rosneft held the last position in production and reserves, yielding less than $0.4(MBOE)/day$ and a proven reserve below $2BBOE$. However, by 2013, it reached the top position, which it maintained consistently until the last year of our analysis, 2021. The peak values for production and reserves were an impressive $5.8(MBOE)/day$ and $44.9BBOE$, respectively.

In contrast, the top three majors, the IOCs Shell, Exxon, and BP, exhibited stable or decreasing production and reserves during the same period. Evidence suggests that NOCs are gaining production and reserve market share over IOCs among these top 10 selections, while the average investments in these two groups show the opposite trend. This is directly linked to the availability of new exploration areas and tenders.

On average, the proven reserves/production ratio for NOCs was 5.85 years, and for IOCs, it was 4.4 years. Among the firms, half are IOCs, and half are NOCs. The average investment values are approximately $2/3$ for IOCs and $1/3$ for NOCs. For production, the total is around $25MBOE/day$, which represents $1/4$ of world production. The values are more balanced for proven reserves, with a 60% to 40% ratio between IOCs and NOCSs, respectively.

These observations indicate some pressure in reserve replacement, requiring either increased investments or reduction in production to maintain a balance in the proven reserves/production ratio. On the other hand, discoveries require substantial investments to develop the oil field, as exemplified by Petrobras' development of the pre-salt field. An exception may occur if the geological features and production costs are reasonably low, as Rosneft in the Siberia basin exemplifies.

Figure 4: Historical Data for Firm and Market Variables



Note: Displays of historical data for key variables across firms, arranged from top left to bottom right. The variables include investment metrics (total year investment, upstream, and exploration percentages), financial aspects (firm volatility and return), and physical performance indicators (daily production, year-end proved reserves, reserve variation, and annual reserve increment).

4.1.2. Game Configuration Analysis

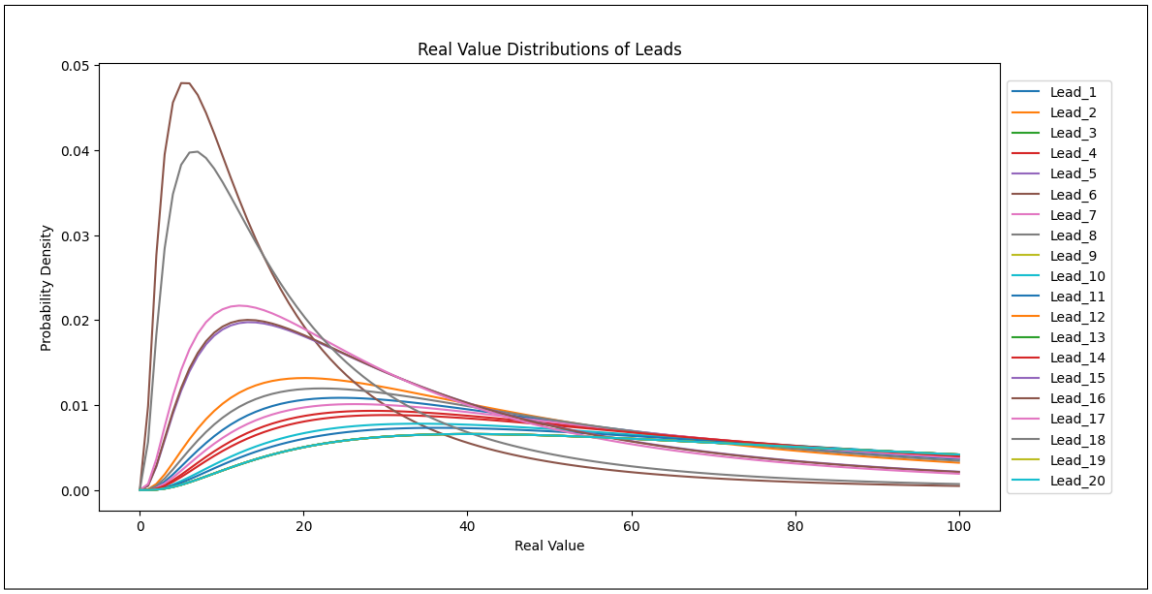
To start simulating the game, it is first necessary to establish the bidding and scenario configurations. These dictionaries outline the dynamics of the game and, consequently, their results. The bidding process is the firms' first contact with the potential assets they can bid on to add to their portfolios, when companies assess opportunities and risks to form their strategic decisions. We have designed the *BidGenerator* class to generate numerous configurations, each associated with a probability curve that represents the uncertainty inherent in real bidding scenarios.

Figure 5 provides an illustrative example of the results of the *BidGenerator* class, showcasing twenty different probability curves. These curves use a log-normal distribution, a type of curve widely adopted in oilfield models due to its ability to represent the 'skewed' nature of lead values in the industry. The log-normal distribution is advantageous because it can model the longer tails where extreme values, although rare, can significantly impact future returns and, therefore, influence decision-making processes to the same degree.

The probability curves aim to ensure that the agents are exposed to a wide range of scenarios during their training. By encountering a variety of potential outcomes, agents are able to gain a better understanding of the environment, increasing the robustness and adaptability of the model. This diversity of scenarios prepare agents to navigate the unpredictability of the game as they learn to make informed decisions under different levels of uncertainty and risk, guided by accumulated experience.

In each round of the game, the firms analyze the generated leads by referencing these probability curves. They evaluate the "true value" of each lead, which serves as a baseline for

Figure 5: Leads Value Distribution for O&G Game



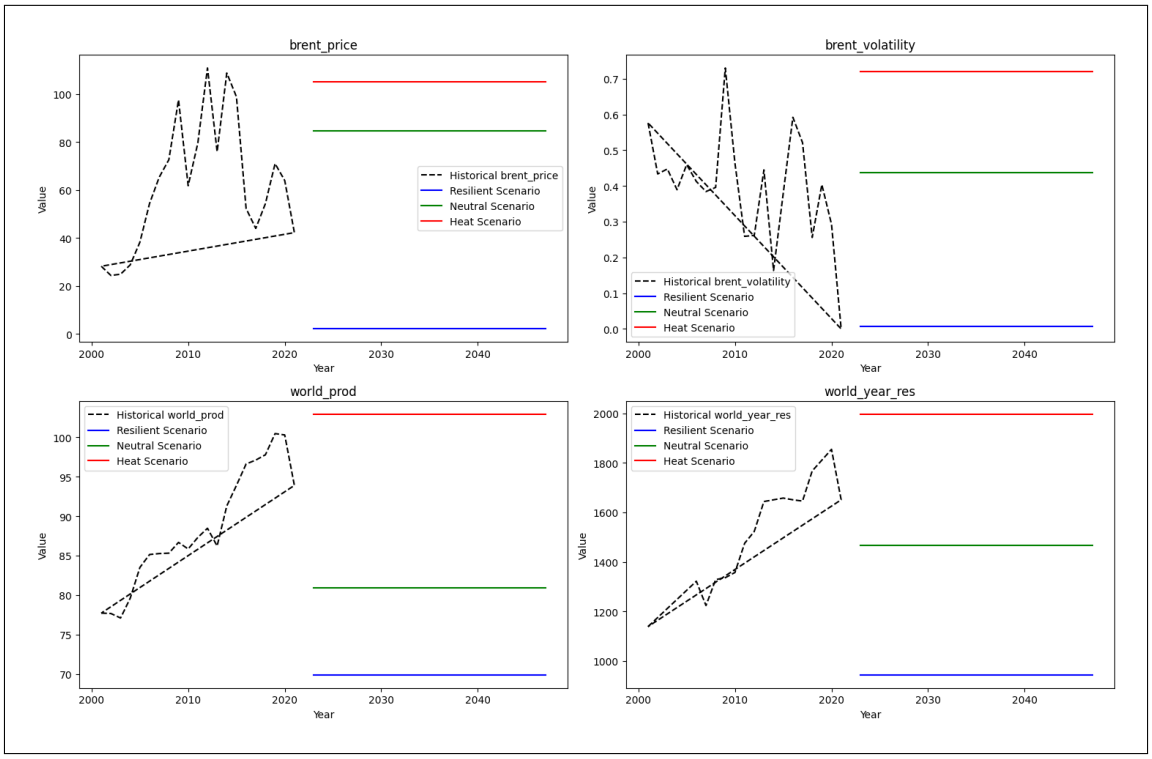
Note: Example of *BidGenerator* output with twenty distinct Log-Normal probability curves.

potential profitability. However, this real value is not constant. It is estimated differently by each company due to their unique risk functions and the individual way they interpret and add market information into their proposals. In other words, each firm risk function is a determining factor in this evaluation process, as it reflects its tolerance for uncertainty and its strategic approach to capital investment. Companies exhibiting greater tolerance for risk may place more aggressive bids on projects with a more significant variation in potential returns. In contrast, more risk-averse companies may prefer projects with lower volatility and more predictable outcomes.

Scenario configurations, generated by the *ScenarioGenerator* class, provide a dynamic backdrop against which companies operate, inputting different economic contexts into the environment. These scenarios consist of probability distributions of various global variables included in the environment, such as oil prices, volatility, and demand fluctuations. Figure 6 shows an example of these configurations, indicating how different market conditions can influence companies' strategic decisions.

As the game progressed, the companies continually evaluated their positions in each round. They compared the evolution of the actual value of the leads with the calculated potential value derived from their risk-adjusted models. This iterative process simulates the real-world decision-making cycle in the oil and gas industry, where companies have to quickly make adaptations to new information in unstable situations. Integrating diverse probability curves and dynamic scenarios ensures that the agents are thoroughly prepared to navigate the game's complexities. By engaging with these sophisticated models, the firms can develop strategies that balance risk and reward, ultimately leading to more informed and effective decision-making processes.

Figure 6: Future Scenarios for O&G Game



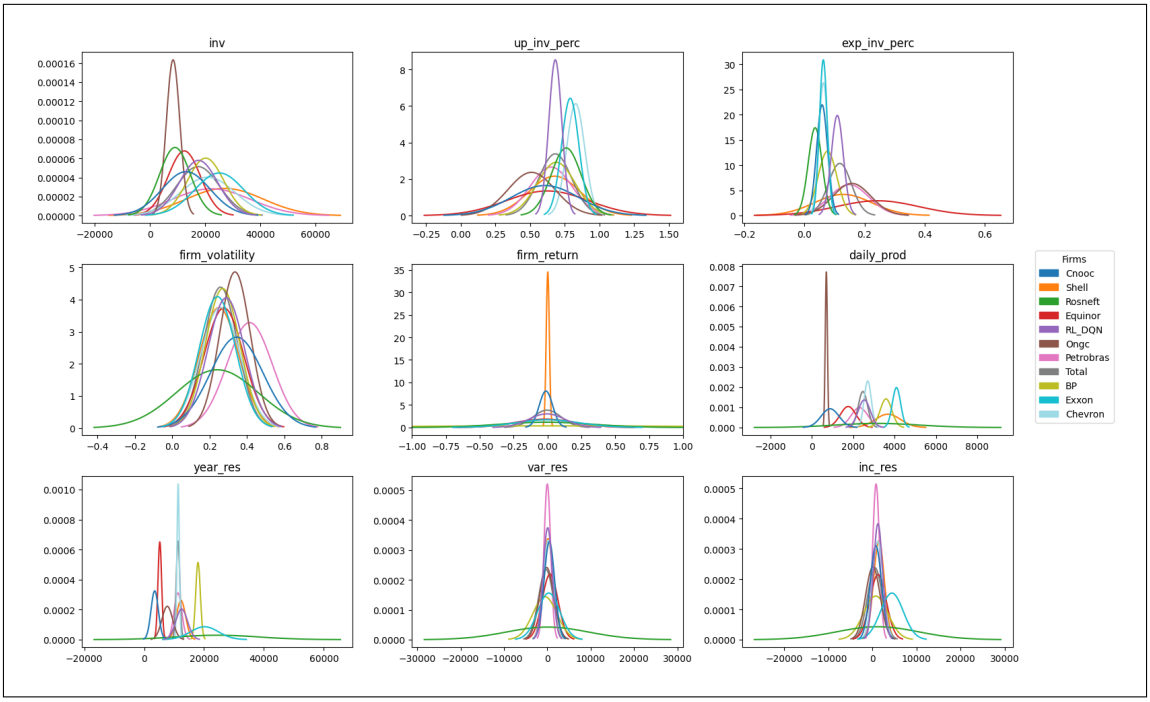
Note: Example of *ScenarioGenerator* output. Displays of both historical and projected values for market variables. The top panels show the Brent price and its volatility, while the bottom panels illustrate global production and reserves measured in billion barrels of oil equivalent (BBOE).

After loading the data, we can obtain the agent profiles representing the characteristics of the companies with the variable that comes from the dataset. This encapsulates profiles of the unique attributes of each firm and passes this information inside the simulation. In Figure 7, we can see these profiles for all firms in our dataset. We use Gaussian distributions that provide a probabilistic representation to calculate the profiles.

Estimating these profiles is only possible by integrating the *Profile* class with the *Agent* class. These classes generate a comprehensive representation of firms, taking advantage of historical data processed in the previous step. The *Profile* class processes the data to capture the main statistical properties and save it appropriately in dictionary format, while the *Agent* class uses these profiles to input them into the companies' decision-making processes in the game environment.

Moreover, these profiles are not a passive representation but actively contribute to the core mechanics of the simulation. The calculation of potential returns, the implementation of risk functions, and the derivation of reward functions all rely heavily on the attributes defined within these profiles. As firms move through the game's different stages, these profiles are integral in determining their strategies and outcomes. For instance, when the environment searches for the winner in the bidding phase, it utilizes the firm's profile to assess its potential success based on calculated risks and projected returns.

Figure 7: Firms Profiles Based on Gaussian Curves



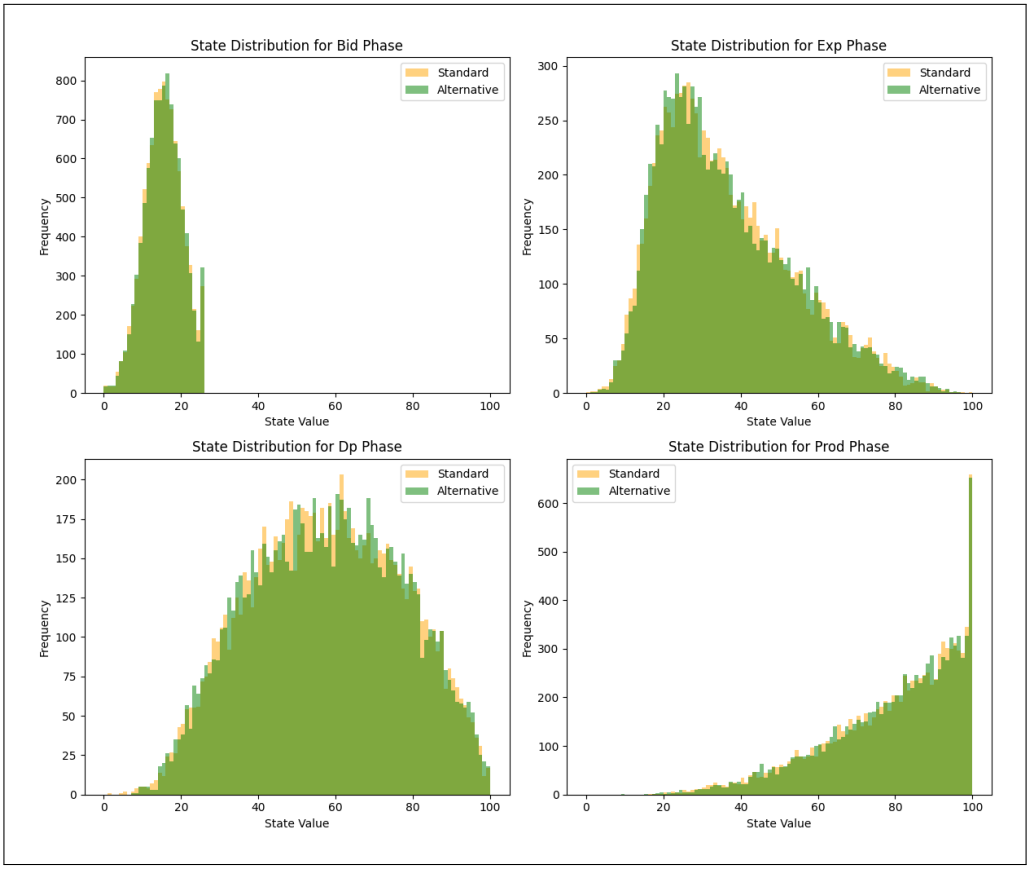
Note: This figure displays Gaussian curves for key variables across firms, arranged from top left to bottom right. The variables include investment metrics (total year investment, upstream, and exploration percentages), financial aspects (firm volatility and return), and physical performance indicators (daily production, year-end proved reserves, reserve variation, and annual reserve increment). Each subplot highlights the distribution and distinctive characteristics of the firms within the simulation.

4.1.3. RL Analysis

After completing these three initial steps, generating a bid, generating scenarios, and calculating the agent's profile, we have the inputs needed to simulate and run the game. Although these outputs can be considered intermediate steps, they provide valuable insights, mainly through the company profiles, enabling comparative analysis between companies. The study of company profiles presents a fertile ground for discussion, as it allows us to examine how other companies are positioned in the market and how their unique attributes influence their strategic decisions.

To perform the experiments correctly, the agent must access the states and actions following the rules established in their respective classes. Figures 8 and 9 show examples of the behavior exhibited by sampled agents, each with different profiles. These figures show the states and actions chosen by the agents in the different phases of the simulation. The firm's behavior results from sampling the state and action spaces. The graphical representations here serve as a valuable tool for analyzing the samples' distribution and assessing whether the system is performing well or as expected.

Figure 8: States by Phases in Simulation with 10k Rounds



Note: The observed states are presented phase by phase, starting from the top-left and moving to the bottom-right. In the bid phase, the states are concentrated in the first quartile of the state space, though they can extend up to half of the available states in other experiments. Following the bid phase are the exploration, development (DP), and production phases, where a more gradual distribution of states across the state space is observed.

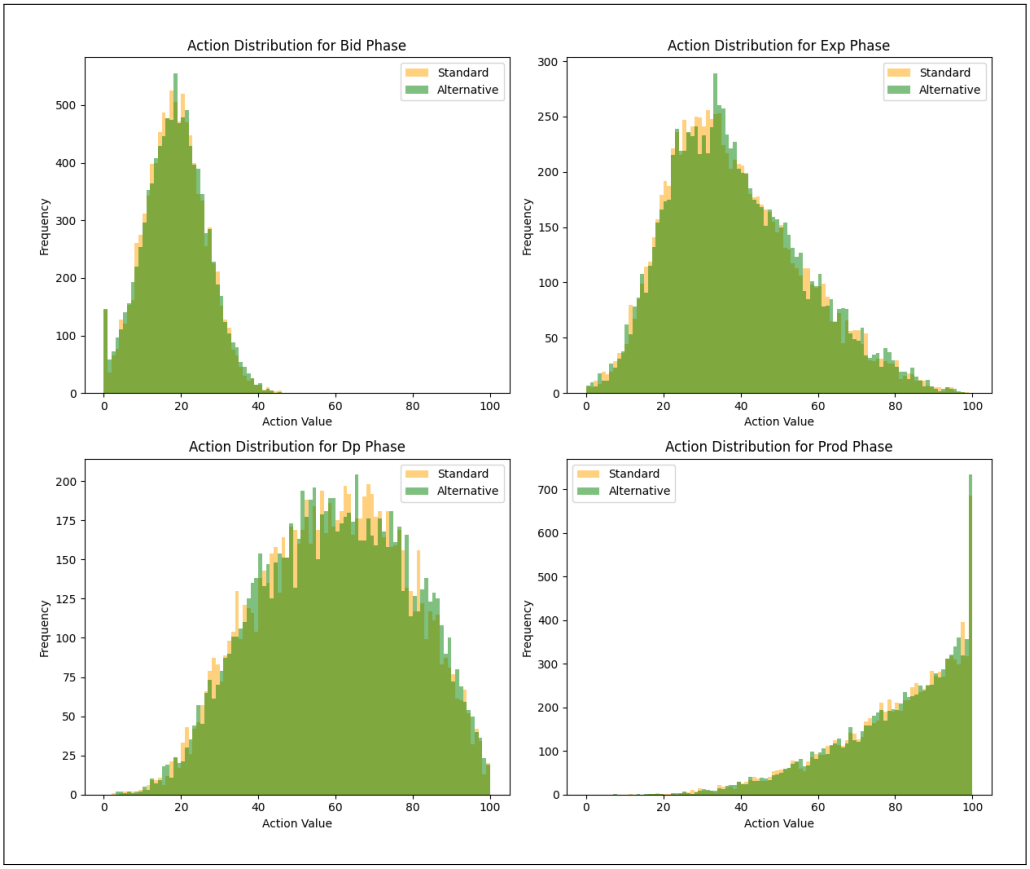
To conclude this section, we present the rewards obtained during the training, as shown in Figure 10 as below. The reward distribution clearly demonstrates an offset in the x-axis and the rewards in the histogram. This occurs because learning agents improve their action choices during the game and are attracted to the most favorable states to maximize their rewards. The multi-modal distributions reflect the variations in bid and market configurations within this experiment. This figure represents just one instance: the game, with dozens of thousands of rounds and experiments, will have a wide range of outcomes. Nevertheless, this example is enough to demonstrate how the RL process proceeds, guiding the agents to optimal strategies.

The following section will examine the policies and results at this work's heart.

4.1.4. Policy Results

As we progress towards evaluating the first results of the reinforcement learning (RL) process, our focus shifts to the analysis of the Q-Table. The Q-Table is a crucial component in RL, as it encapsulates the learned values that the agent associates with each state-action pair. These values, known as Q-values, guide the agent's decision-making process by indicating the expected utility (future reward) of taking a specific action in a given state.

Figure 9: Actions by Phases in Simulation with 10k Rounds



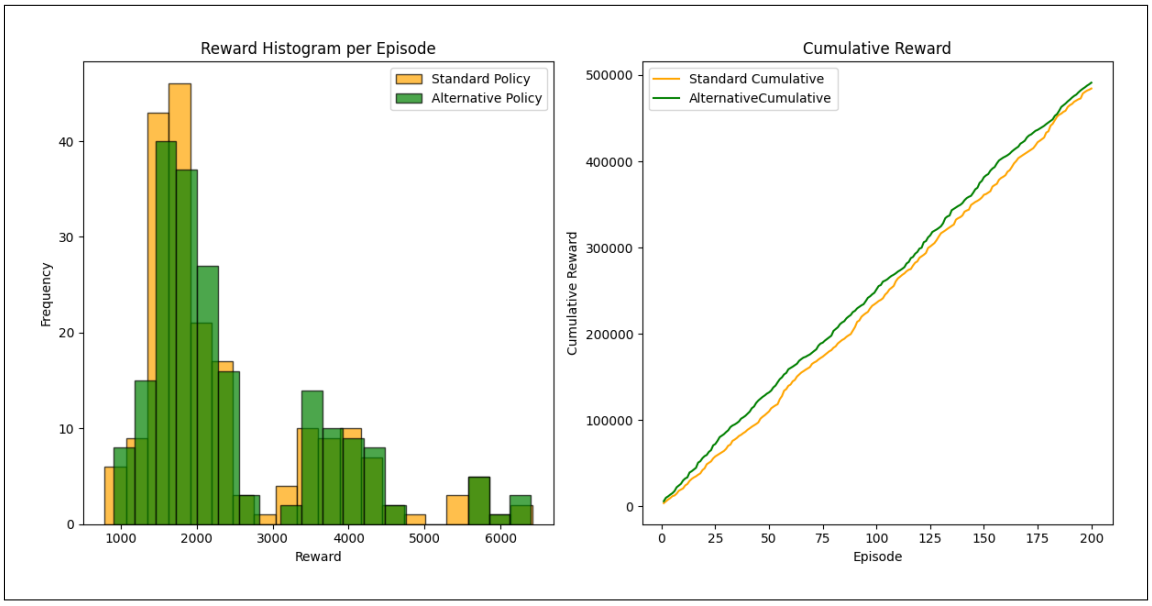
Note: The observed actions are presented phase by phase, starting from the top-left and moving to the bottom-right. In the bid phase, the states are concentrated in the first quartile of the state space, though they can extend up to half of the available states in other experiments. Following the bid phase are the exploration, development (DP), and production phases, where a more gradual distribution of actions across the action space is observed.

To visually interpret the Q-Table, we represent it as a Heat Map in Figure 11 , where the color's intensity corresponds to the magnitude of the associated Q-value. This visualization provides a clear and intuitive understanding of the agent's learning progress and the efficacy of its decision-making strategy. By analyzing the Heat Map, one can identify patterns, such as which states lead to higher rewards and how the agent prioritizes different actions in various scenarios. This analysis is pivotal for understanding the agent's learning process dynamics and identifying potential areas for further optimization.

While the intermediate outputs from the bidding, scenario, and agent profiling stages set the foundation for the game, the Q-Table analysis offers a window into the agent's evolving strategy. It offers valuable insights into the efficacy of the reinforcement learning approach and lays the groundwork for further refinements and advancements.

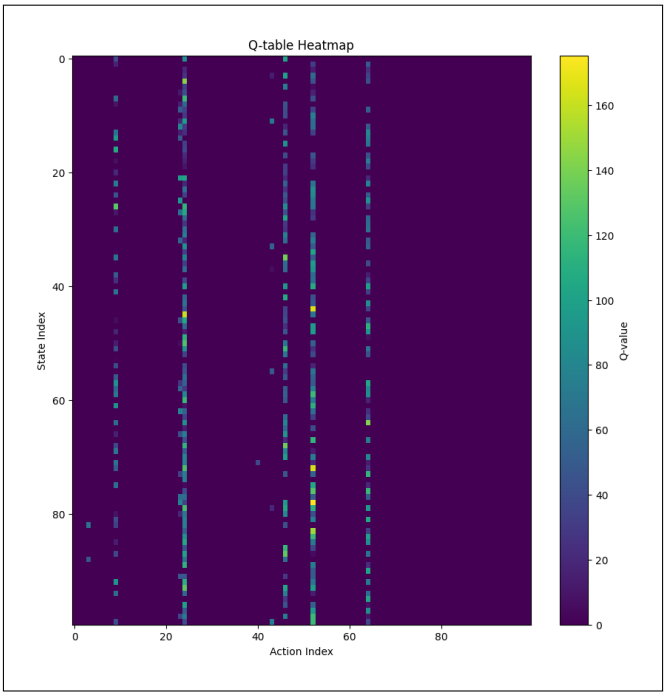
The main results of these experiments are summarized in Figure 12. This figure presents the actions recommended by the policy to maximize returns across a range of bidding scenarios, environmental conditions, and agent profiles. The analysis compares the decisions made by trained and untrained agents, highlighting the policy's impact on decision-making processes. In

Figure 10: Reward and Cumulative Reward by Episode



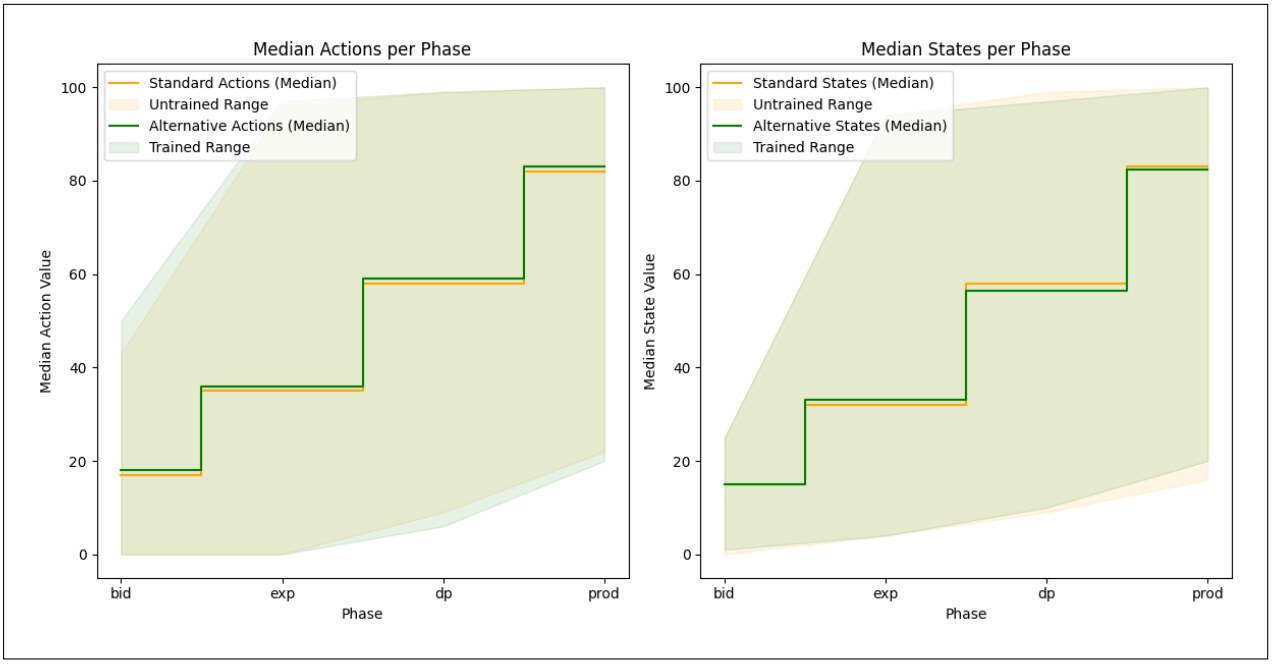
Note: In this picture, we compare two strategies in given bid competition and market conditions where the agents learn using the Q-learning method. In the right-side figure, we observe a multi-mode pattern in the reward distribution histogram with a slight offset in favor of alternative strategies. The left-side figure shows that the learning agents have performed better than the untrained or standard strategies since the beginning.

Figure 11: Q-Table Heat Map for Symmetric Action-State Space



Note: Heat Map of the Q-table with a 100x100 state-action space. Each cell corresponds to the expected cumulative reward for a given state-action pair, learned through Q-Learning with an exploration rate of $\epsilon = 0.2$. The color intensity indicates the magnitude of the Q-values, where lighter colors correspond to higher Q-values (more favorable actions), and darker colors represent lower Q-values.

Figure 12: Policy by Phase for the Alternative Versus Standard Strategy



Note: The left plot shows the median actions and the right plot displays the median states. Shaded areas represent the observed space, with the x and y axes depicting action and state spaces. Strategy differences are minimal due to the limited rounds, decreasing further with more rounds. The best strategy identified involves early and increased investment, while the development phase has a lower median in information quality compared to the optimal strategy.

addition, we employed the stepped policy as a ground truth, where firms adjust their investment strategies in alignment with decreasing risk levels, providing a benchmark for evaluating the effectiveness of the learned policies.

The DQN method exploits network nodes that are challenging to interpret, as they are not correlated or do not correspond to visual representations of real-world features. Consequently, to avoid ambiguity in the direct analysis, we have chosen not to present the network values through a visual representation.

The game's logic is complete for the Q-Learning approach, with the simulations designed to manage the training and evaluation of agents in a multi-phase environment. The simulation is initialized with the self-play training, where each firm participates in simulations, choosing actions based on an epsilon-greedy policy and updating the Q-Tables by self-playing. During each simulation, the firms interact with the environment, and the Q-tables used to represent the value of different state-action pairs are updated based on the rewards received. Over time, the epsilon value decays, gradually shifting the firm's focus from exploring new actions to exploiting the best strategies it has learned. This allows the firm to fine-tune its decision-making process for the next training steps.

Once the self-play training is complete, we can evaluate the performance of the trained firms in a competition. Here, the firms use their trained Q-tables to make decisions, with exploration

minimized so they can focus on exploiting their learned strategies. The firms then interact with the environment across different phases, such as bidding, exploration, and production, and the rewards they collect are tracked. We have introduced some variability based on the dataset into the rewards to represent real-world uncertainty. The competition results reveal which firms performed the best in each phase, offering valuable insights into strategies that yielded the best performances during training. We display in Figure in the appendix is the example of self-play Q-Tables for all firms in our set of agents.

Based on these competition results, we use this outcome to refine the firms' strategies. By conducting rollouts with the recorded best actions, firms can simulate games employing the top actions identified in the competition, helping them improve their policies. This setup creates a feedback loop where firms continuously improve their strategies based on their learning and insights from competing with others, fostering a dynamic and iterative approach to policy development.

The agent training process initializes with a predefined environment, policy, and configuration, which facilitates the training over multiple episodes and simulations. The core of the training process involves updating the agents' policies based on the experiences they gather during interactions with the environment. The training method integrates the best actions observed from competitions in simulations, allowing firms to refine their strategies. This method iterates over selected firms and episodes, ensuring that the Q-values in the policy are updated based on the rewards and actions taken in each state regarding the features of the bid and scenario. In this step, we introduce a mechanism to generate multiple bid configurations, update the environment, and train the agents under each bid. This method works in tandem with the generation of scenarios. By doing so, the training ensures that agents are exposed to a wide range of conditions, simulating different real-world possibilities and allowing them to develop robust strategies. This layered approach to training across both bids and scenarios aims to improve the agents' performance in diverse situations. Table 2 summarizes the bid, scenario, and firm configurations during one of the experiments. The appendix Figure shows the result of one entire game round, the rewards for each phase, and the total.

Table 2: Summary of Bids, Scenarios, and Firms Selected for Training

Statistic	**Mean**	**Std Dev)**
True Value	1000.00	100.00
Oil Price	60.00	5.00
Oil Price Variance	5.00	1.00
Firms Selected	7.00	4.00
Bids	100	N/A
Scenarios per Bid	100	N/A
Total Iterations	10000	N/A

Note: True Value is the value of the lead in the bid. Oil price and variance are obtained in profiles for global variables. Firm selection varies between 2 and 11 firms, including the learn agent. Bids and Scenarios are the number of iterations we used for the final training.

The detailed training results, including the bid and scenario information and the resulting trained policies for each combination, are used during the strategy evaluation. These results are then accessible, providing a comprehensive overview of how the agents performed under different conditions. This design supports deeper analysis and evaluation of the training process, ensuring that the agents' strategies can be assessed in the context of varying bids and scenarios.

This section presented examples of all classes used to train the agents. The visualization investigates the best way to model the problem and how to implement it in the codes. The final results are presented in the next section, where we briefly discuss the performance and refinement of the models and mainly evaluate the strategies found at the end of the training. Let us move on to these discussions and confront them with the research hypothesis.

4.2. Discussions

4.2.1. Model Evaluation

In this RL system implemented for the O&G context, we offer a structured and modular approach to simulating decision-making processes for this industry. Our experiments indicate that this system captures the complexity of real-world scenarios by using a series of interconnected classes such as the environment and agent models, reward mechanisms, and policy strategies. The build of this software was based on a modular method to ensure that each component could be independently developed and tested, enhancing the overall robustness of the model. Additionally, by grounding simulations in historical data, the system can align the RL agents' learning processes with actual industry dynamics, improving the reliability and relevance of the generated results. However, the system also presents challenges, particularly in balancing the simplicity of discretized action and state spaces with the need for detailed, continuous decision-making in complex environments. Future work may explore more sophisticated reward structures, continuous state spaces, and comparative algorithm analyses to enhance the system's applicability and performance. In the following lines, we discuss the main aspects of the model, such as its strengths and weaknesses, to understand how the results and investigations are reliable and robust.

One of the key strengths of this RL system and its model lies in its comprehensive approach to modeling the environment, agents, and reward mechanisms. Employing distinct classes such as LoadDataBase, Profile, BidGenerator, ScenarioGenerator, Agent, Action, State, Reward, Environment, and Policy ensures a modular system, which is both managable and extensible, allowing the independent development and testing of each component.

The general model, including trained and untrained agents in the evaluation process, allows for a comparative analysis of policy effectiveness. This comparative approach is valuable for understanding the impact of different training strategies on agent performance, providing a clear metric for assessing the success of the RL system. The visualization tools, such as Q-Table heatmaps and action-state distribution plots, further aid in interpreting the outcomes, making the results more accessible and understandable.

Despite its strengths, the system exhibits specific weaknesses that could affect the reliability and generalizability of the results. One notable concern is the potential oversimplification of the action and state spaces. While discretazing action and state spaces simplify the computational process, they may need to fully capture the complexity of decision-making in the oil and gas industry, where continuous variables and more nuanced decisions play a significant role. The current discretization might lead to sub-optimal policies that do not generalize well to more complex, real-world scenarios.

Another potential weakness is related to the reward structure. While logically sound, the reward calculation may benefit from a more sophisticated approach that considers long-term consequences and opportunity costs associated with different actions. The current reward function is primarily focused on immediate outcomes, which may lead to short-sighted policies that

do not adequately account for the long-term sustainability and profitability of decisions made by the agents.

Furthermore, the system's reliance on specific parameters such as the discount factor (γ), learning rate (α), and exploration rate (ϵ) raises concerns about parameter sensitivity. The performance of the RL agent is heavily influenced by the proper tuning of these hyper-parameters. Inadequate tuning could lead to convergence on sub-optimal policies or excessive exploration that hinders effective learning. This underscores the importance of conducting extensive parameter sweeps and sensitivity analyses to guarantee the robustness of the learned policies.

The reliability of the results from this RL system depends on how accurate the environment and agent models are and how well the training process is done. Using historical data to guide the bidding and scenario generation processes makes the simulations more realistic, which increases the chance that the agent's learned policies will work in real-world situations. The system's ability to generalize to new and unseen scenarios is contingent upon the training data being diverse and representative. Therefore, the RL agent may perform well outside the simulated environment if the training scenarios cover a wide range of real-world conditions.

To boost generalizability, we could explore advanced techniques, like continuous action spaces, or sophisticated exploration strategies such as softmax or Thompson sampling which could prove beneficial. Additionally, incorporating mechanisms for the agent to adapt its policies based on new information or changing market conditions could further improve the system's ability at dealing with dynamic situations.

4.2.2. Fine-Tuning and Performance

Fine-tuning is a process that involves systematically tuning the hyper-parameters of a reinforcement learning model, in our case specifically α (learning rate), γ (discount factor), and ϵ (exploration rate). The process starts with a grid search, where various combinations of these hyper-parameters are tested. For each combination, the agent is trained through interaction with the environment, and its performance is evaluated using three key metrics: cumulative reward, success rate, and Average Squared Error (ASE). The results of these evaluations are stored for later analysis. After the grid search, the results are sorted, and statistical values such as the worst, first quartile, median, third quartile, and best results are selected. This allows the model to identify the best-performing configuration and provide a performance summary across various hyper-parameter settings.

The following we have tables summarize the results of fine-tuning the hyper-parameters (alpha, gamma, epsilon) for the Q-Learning as we see in the Table 3, SARSA that can be seen in Table 4, and DQN that is showed in Table 5. The tables include different sets of hyper-parameters and their corresponding performance metrics, such as cumulative reward, the number of successful episodes, or action selection efficiency. The cumulative reward is the total reward accumulated over all episodes. The success rate metric refers to the percentage of successful outcomes or tasks the agent completes. During simulations, it measures how often the agent achieves its predefined goals or desired states. A higher success rate indicates that

the agent consistently performs well and makes correct decisions. The action selection efficiency metric (ASE) reflects how efficiently the agent selects optimal or near-optimal actions during interactions with the environment. It is typically measured by comparing the agent's chosen actions with the best actions based on learned policies. A higher ASE suggests that the agent effectively learns and applies policy to make informed decisions, reducing unnecessary exploratory or sub-optimal actions.

In Table 3, the results show that different combinations of hyper-parameters lead to varying performances in Q-Learning. The combination of $\alpha = 0.10$, $\gamma = 0.7$, and $\epsilon = 0.5$ stood out, with the highest cumulative reward of 8831 and a 100% success rate, although with a high ASE of 75720. Another interesting setup was $\alpha = 0.05$, $\gamma = 0.1$, and $\epsilon = 0.7$, which also achieved a 100% success rate but with a much lower cumulative reward of 1681. The other combinations showed more modest results, with cumulative rewards ranging from 269 to 831 and lower success rates. This highlights how tuning the hyper-parameters can make all the difference in the outcome, which is a step that can be neglected during the development of RL models.

Table 3: Performance for Selected Hyper-Parameter Settings in Q-Learning

Hyperparameter	**Cumulative Reward**	**Success Rate (%)**	**ASE**
$\alpha = 0.50 \mid \gamma = 0.7 \mid \epsilon = 0.5$	269	58	281
$\alpha = 0.01 \mid \gamma = 0.7 \mid \epsilon = 0.1$	566	73	493
$\alpha = 0.05 \mid \gamma = 0.3 \mid \epsilon = 0.1$	831	93	680
$\alpha = 0.05 \mid \gamma = 0.1 \mid \epsilon = 0.7$	1681	100	2626
$\alpha* = 0.10 \mid \gamma* = 0.7 \mid \epsilon* = 0.5$	8831	100	75720

* Best results for these parameters.

Table 4 below shows the outcomes for various SARSA hyper-parameters. We notice modest cumulative reward and success rate performance at lower alpha, gamma, and epsilon values. For example, with $\alpha = 0.05$, $\gamma = 0.50$, and $\epsilon = 0.7$, we see a reward of 196 and success at 51%. Performance improves as these values adjust, but the standout result is the configuration $\alpha = 0.05$, $\gamma = 0.99$, $\epsilon = 0.7$, delivering the best reward of 3775 and a perfect 100%

Table 4: Performance for Selected Hyper-Parameter Settings in SARSA

Hyperparameter	**Cumulative Reward**	**Success Rate (%)**	**ASE**
$\alpha = 0.05 \mid \gamma = 0.50 \mid \epsilon = 0.7$	196	51	268
$\alpha = 0.10 \mid \gamma = 0.30 \mid \epsilon = 0.9$	315	58	279
$\alpha = 0.10 \mid \gamma = 0.10 \mid \epsilon = 0.5$	378	63	347
$\alpha = 0.01 \mid \gamma = 0.50 \mid \epsilon = 0.5$	507	69	579
$\alpha* = 0.05 \mid \gamma* = 0.99 \mid \epsilon* = 0.7$	3775	100	13439

* Best results for these parameters.

Ultimately, Table 5 below presents the results for DQN hyper-parameters fine-tuning. Initially, we observe a gradual increase in cumulative rewards and success rates as the hyper-parameters adjust. For instance, with $\alpha = 0.20$, $\gamma = 0.99$, and $\epsilon = 0.1$, the cumulative reward is just 148 with a success rate of 53%. As we experiment with different values, performance

improves, especially with $\alpha = 0.10$, $\gamma = 0.99$, and $\epsilon = 0.1$, achieving a 76% success rate. However, the best configuration by far is $\alpha = 0.01$, $\gamma = 0.50$, and $\epsilon = 0.9$, which hits a 100% success rate and a much higher reward of 2681.

Table 5: Performance for Selected Hyper-Parameter Settings in DQN

Hyperparameter	**Cumulative Reward**	**Success Rate (%)**	**ASE**
$\alpha = 0.20 \mid \gamma = 0.99 \mid \epsilon = 0.1$	148	53	204
$\alpha = 0.50 \mid \gamma = 0.99 \mid \epsilon = 0.7$	306	64	338
$\alpha = 0.05 \mid \gamma = 0.50 \mid \epsilon = 0.9$	362	68	238
$\alpha = 0.10 \mid \gamma = 0.99 \mid \epsilon = 0.1$	546	76	374
$\alpha* = 0.01 \mid \gamma* = 0.50 \mid \epsilon* = 0.9$	2681	100	6684

* Best results for these parameters.

The DQN model outperformed both Q-Learning and SARSA, achieving the highest cumulative reward and success rate, and lowest ASE values. Q-Learning performed well, particularly in cumulative reward, while SARSA lagged slightly behind the overall performance. Overall, DQN offered the optimal balance between learning efficiency and robustness outcomes. The fine-tuning tables illustrate the impact of hyper-parameters on Q-Learning algorithm performance. In contrast, the algorithm comparison table facilitates a direct comparison of SARSA, Q-Learning, and DQN, highlighting the trade-offs inherent in each methods. In general, DQN offers key advantages such as using neural networks for Q-value approximation of each state-action pair, enabling it to deal with higher dimensional spaces and complex environments. Furthermore, its function approximation capability provides better generalization across a broader range of states and actions. Additionally, the experience of replay through replay buffers and random batch sampling during training improves data efficiency and stability learning.

Looking at Table 6, we can quickly compare how SARSA, Q-Learning, and DQN perform in reinforcement learning tasks. These algorithms were tested for their average cumulative reward, how fast they converged, and their success rate. The table shows that all three algorithms perform similarly in terms of success rate and convergence speed, with each reaching 100% success and converging after 1000 episodes. However, their average cumulative rewards are slightly different, with SARSA having a slight edge, followed closely by Q-Learning and DQN.

Table 6: Performance for SARSA, Q-Learning, and DQN

Metric	**SARSA**	**Q-Learning**	**DQN**
Average Cumulative Reward	87271	86981	86231
Convergence Speed (Episodes)	1000	1000	1000
Success Rate (%)	100	100	100
Robustness to Scenarios	Moderate	Moderate	Moderate

All three algorithms are marked as "moderate" regarding robustness under varied or unseen scenarios. Despite SARSA's slight advantage in cumulative reward, the overall performance of the three methods is quite close, with none outperforming the others in these tests.

4.2.3. Strategy Analysis

Historically, firms that operate in markets with exploratory characteristics face a high-risk initial phase. This phase presents a dilemma with low monetary returns and assets of high potential value, such as the oil and gas industry. The decision to apply more or fewer resources is crucial, given the high risk of project discontinuity and the need for better resources. In this exploratory phase, the main costs are related to information assets, adding to the gravity of the situation.

For the oil industry, a crucial factor is the time it takes to obtain information, whether it's environmental licensing time, data acquisition operations, or information processing. Seismic and well data can take one to four years to be available for decision-making. However, several investment decisions in production development occur sequentially, and new information cannot wait. Poor-quality information can significantly impact development phase investments, notably higher since production units and systems form the bulk of the entire chain and can reach 80% of the total investment.

In addition, it is not uncommon for data to be redundant in the same area, where firms repeatedly acquire higher-quality data as the risk decreases. This leads to issues that can result in inefficiencies in corporate investments. The first point is the life cycle of the information, which becomes very short, reducing the asset's usefulness in each new investment. Obtaining this information, in the case of offshore exploration, uses vessels that occupy areas for months, emitting greenhouse gases. Avoiding repeated surveys and reducing environmental issues also reduces costs in advanced phases, beyond supporting better decisions in production development in the most intensive investment phase.

With this general economic analysis and methodologies refined, we focus on the strategies and results from the experiments. For that, we have a metric that compares the higher return strategy by phase, allowing us to analyze whether our hypothesis research is compatible with the learned agents. The first metric concerns the existence of alternative strategies that anticipate investment in the early phases of the O&GI. In this case, we define the ratio of times the alternative strategy provides early investment and the outcome of return by the standard or untrained strategy. For the second hypothesis, we must check these strategies in the context of various bids, scenarios, and agent configurations playing the game. For that, we run and save configuration game parameters and split the results for the domain to see if the strategies keep the behavior across the entire set of parameters.

In this part of the work, we have the evaluation pipeline generating several important metrics related to the performance of strategies that anticipate actions in earlier phases of the reinforcement learning process. Following a natural sequence, we calculate the success rate, a metric that reflects the percentage of episodes where the cumulative reward exceeds a pre-defined threshold. In the context of strategies that prioritize earlier actions (e.g., making decisions on bidding, exploration, development, or production earlier in the process), the success rate helps to assess how often these early actions with higher values of investments lead to favorable outcomes compared in both standard or alternative policies. A higher success rate indicates that anticipating actions early benefits a more significant proportion of bids, scenarios, and competition.

The strategies here are classified as Standard and Alternative, which means they are untrained and trained in the RL sense. The standard strategies allow the anticipated investment but have constraints in the investments by phase, following a staircase design with the following investment always being done and more significant than before action value. As an alternative, we do not have constraints that are possible. The action in some phases is zero, jumping the investment in such phase in the case that this is found by the agent, whether through e-greedy or not. For that, we observe, in many cases, some convergence between the two possible strategies. The idea is not to prove that the standard strategy is wrong or worse than the other but to propose alternatives that can result in better rewards in the long term.

Now, we can measure the average squared error (ASE), taking the average deviation between the predicted rewards, used as a benchmark, and the actual rewards obtained during the episodes. In the case of early action strategies (ES), this metric is helpful in determining how closely the actual performance aligns with the expected outcomes. A lower ASE indicates that the anticipated actions are more consistent with predicted values and that the strategy is more reliable.

Additionally, we have monitored the convergence speed, which stores the number of episodes required for the policy to stabilize or converge towards optimal behavior. Convergence speed evaluates whether such early decisions help or hinder the learning process in strategies where actions are taken earlier. A faster convergence suggests that early actions provides more precise an clearer signals to the agent, allowing it to quickly learn optimal behavior.

Another metric tested was the Robustness of Scenarios. This qualitative metric evaluates the performance of the early action strategy across different scenarios or environments. Robustness is particularly important in reinforcement learning because it indicates whether the strategy that anticipates actions earlier is effective across a wide variety of conditions (e.g., different market conditions, competitor behaviors, or resource availability). A strategy that maintains high performance across diverse scenarios is considered more robust. This fits well with our purpose in the second hypothesis assertion. However, the most critical metric is early action metrics, such as the Early Success rate (ES). These refer to specific performance evaluations of strategies prioritizing actions in earlier phases (bidding, exploration, development, production).

Tables 7, 8, and 9 summarize the experiment's results. Figures 15, 16, and 17 in the appendix allow us to visualize these metrics, showing whether taking actions earlier in the process leads to better cumulative rewards or quicker learning. These metrics help compare strategies that take action earlier with those that wait for more information before making decisions.

Competition level is a feature that significantly affects both the early investment metric and the efficiency of the training. The results in Table 7 and Figure 15 have higher values for ES for the alternative strategy for many firms. Running the game many times shows that the average number of firms changes. However, the pattern of decreases in the anticipated investment performance with the increment of competition level has been found in most runs. On the other hand, the results for the alternative strategy have a higher average reward than the

Table 7: Summary of Early Strategy Metrics Using DQN by Competition

Firms	ES Alternative	ES Standard	Avg Rwd Alternative	Avg Rwd Standard
2	100.00	100.00	32.40	78.06
3	95.73	95.73	225.84	309.26
4	98.08	96.15	149.75	218.46
5	97.10	97.10	209.34	241.43
6	97.75	97.75	342.27	274.44
7	96.55	97.41	187.15	163.28
8	99.04	97.12	227.88	208.24
9	95.31	100.00	238.74	188.34
10	100.00	97.09	171.15	212.00

Note: ES is short for Early Success Rate and Rwd for Reward. The column firms correspond to the competition level.

standard with the increment of the competition. This observation is related to the fact there are more firms playing the game and collaborating more with the agents' training.

The second feature to analyze is the scenario with a slight but crescent pattern for ES over the categories. Table 8 and Figure 16 show these results. The average reward by scenario has a more intense difference between categories. The alternative and standard strategy for the scenario and bid features do not differ significantly because the grouping mode has few options. To observe more pronounced differences, we need to consider more categories. However, this introduces a trade-off: as the complexity of numerous categories increase, a macro interpretation becomes less straighforward, which highlights an option for potential improvements in future analysis.

Table 8: Summary of Early Strategy Metrics Using DQN by Scenario

Scenario	ES Alternative	ES Standard	Avg Rwd Alternative	Avg Rwd Standard
Low	96.18	96.18	167.78	167.78
Medium	97.52	97.52	230.95	230.95
High	97.67	97.67	233.04	233.04

Note: ES is short for Early Success Rate and Rwd for Reward. The Low, Medium, and High in this context represent the heat, normal, or resilient market, which translates the level of demand, price, and volatility in the oil market.

The bid features have minimal impact on the categories, with the anticipated investment metric demonstrating favorable outcomes. In this investigation, it's worth remembering that this metric with high values implies the best outcomes for the actions that prioritize the early phases of the game. Even though the bid and scenario outcomes have a negligible impact on the different categories, this metric consistently exhibits high values across all experiments. Based on this observation of this simplified model, we can provide evidence that the presence of an alternative strategy shifts the investment towards the initial phases in the upstream oil and gas industry.

Table 9: Summary of Early Strategy Metrics Using DQN for Bid Feature

True Value	ES Alternative	ES Standard	Avg Rwd Alternative	Avg Rwd Standard
Low	97.62	97.62	187.75	187.75
Medium	97.08	97.08	249.63	249.63
High	97.33	97.33	221.74	221.74

Note: ES is short for Early Success Rate and Rwd for Reward. In this context, the Low, Medium, and High represent the Bid with big, medium, and small leads in a log-normal scale.

4.2.4. Future Research Directions

When we look at the current reinforcement learning (RL) framework, we can see a few promising research paths that could improve how robust and useful it is, especially in economic and industrial settings. One possible way to build on this work would be to look at different reward structures and how they affect how the agent behaves. By tweaking the reward function, researchers can see how different incentive schemes affect decision-making processes, which could lead to more efficient or optimal strategies. This approach is instrumental in complex domains like oil and gas exploration, where small changes in reward mechanisms can greatly impact the results, given the many factors at play.

Furthermore, incorporating continuous variables or higher-dimensional representations within the state and action spaces could facilitate the creation of a more sophisticated and realistic environment for agents to operate within. In oil and gas exploration, decisions are contingent upon a complex network of interrelated variables, including resource availability, market dynamics, and competitor behavior. The simulation of more complex environments with greater state-action space complexity would better equip RL models to reflect the intricacies of real-world decision-making, thereby increasing the model's accuracy.

A further crucial area of investigation is the comparative analysis of RL algorithms across varying conditions. Although the current system relies primarily on Q-Learning, extending the investigation to encompass more sophisticated techniques, such as Actor-Critic models, Proximal Policy Optimization (PPO), or Asynchronous Advantage Actor-Critic (A3C), could provide deeper insights into the comparative strengths and weaknesses of different algorithms. The A3C leverages parallel learning with multiple agents, which allows a broader and faster exploration of the state-action space. The A3C has shown strong performance, often outperforming DQN in various applications within complex and dynamic environments. However, A3C also depends on the characteristics of the environment, and trade-offs between performance and complexity. Therefore, exploring diverse methodologies could enhance our comprehension of deep reinforcement learning techniques and their applications within the oil and gas industry. This analysis would be precious in understanding how these methods perform in environments of varying complexity or with differing amounts of training data. This is a crucial consideration for scalability.

Conclusions

This master's thesis was designed to meet two distinct objectives. One is the academic contribution to investment decisions with RL techniques, while the other is to build a reliable, flexible, and adaptable algorithm for a real firm intern process to optimize strategy and return over the information assets portfolio. Looking more closely at the general and specific goals, we can look into the research hypotheses to assess whether the algorithm developed and the results achieved the initially set objectives.

In the literature review, we looked for the primary references to help connect different research areas, such as corporate investments, valuation, and reinforcement learning. It is worth highlighting that even when covering all topics in theoretical terms, the biggest challenge in this work was implementing the codes. In this case, it is recommended that practical courses be taken on online platforms in addition to the references added to this thesis.

In our investigation to evaluate the possible effects of investment strategies on informational assets in advance, we developed the RL system by simulating a four-phase game in which companies make investment decisions based on the quality of information. The consequences of these decisions are analyzed through the return function. The models are based on profiles and scenarios calculated with real data from firms and oil and gas industry markets. The data used to form the firm profiles and generate the scenarios provides a database built with a combination of data from different sources, as described in the section corresponding to this topic. Given the work carried out in treating this data and the empirical analysis, it is a product.

We apply three techniques in an environment with discretized states and actions to define the optimized policy in the firms' decisions. In this modeling, the policy update was carried out via Q-Learning, SARSA, and DQN, each with advantages and disadvantages. DQN stands out over the others in terms of scalability, given that for discrete spaces without approximation functions, the use of neural networks allows the increase of the space of states and actions without increasing the computational cost, given that the network parameters are fixed to the priory. We observed that the three techniques presented similar results in terms of optimization and that they can be used satisfactorily for the objectives defined here. The fine-tuning showed that the hyper-parameters vary between the techniques used and should be addressed since different hyper-parameters can lead to diametrically opposite results for the same technique. The final analysis's outcomes were done with the DQN configurations with increased neurons in the first and second layers, each one followed by batch normalization for input and hidden layers, a Leaky-Relu activation function, and a dropout layer before the output. The code specifications can be shared if requested.

About our two research hypotheses, it was found that agents trained by RL obtained superior returns, both in the total analysis and segmented by feature. Based on the results and discussions held, it can be seen that the level of competition in the bids is the factor with the most significant weight in the benefits of advanced investments. This result converges with the logic that fewer firms in the bid increase the bid's chances of winning and that the anticipated investment will be fully utilized. On the other hand, the variation in market conditions obtained by modeling the scenarios confirms that for heated scenarios, price, volatility, and demand for oil on the rise increase returns but without generating a significant difference in the benefit of advanced investment in information assets. Concerning the bid, the effects are similar, and it should be noted that for both scenario and bid, the ES values, the metric of gains in anticipating the purchase of data, have high values. It can be inferred that market and bid characteristics have this more minor effect on ES because they are expected to all firms and do not differentially affect returns throughout the training.

The developed database and RL system can be improved to become an analysis tool in real oil and gas industry cases. This improvement can be carried out both in the modeling of base classes, such as states, actions, reward, and environment, to adapt to different approaches or include new agents, such as the regulatory agent, and in techniques, including approximation functions such as the Asynchronous Advantage Actor Critic (A3C) and Direct Preference Optimization(DPO), which has been shown prominently for RL-scaled systems.

In conclusion, we can assert that RL has a promising path within the economic decision-making process, especially in complex analyses involving multiple agents where classical valuation techniques present some limitations. This research contributes to this field of study while fulfilling the technical function of producing a methodology for use in a business environment.

References

[1] Bratvold, R. and Begg, S., (2010), "Making good decisions, *Society of Petroleum Engineers*.

[2] Suslick, S., Schiozer, D., and Rodriguez, M., (2009), "Uncertainty and risk analysis in petroleum exploration and production.*Terrae*, vol 6, n 1, pp. 30–41.

[3] J. Plate, *et al.*, (2004), "Occlusion Culling for Sub-Surface Models in Geo-Scientific Applications.", 10.2312/VisSym/VisSym04/267-272, *Inproceedings*, vol 9, pp. 267–272, 351.

[4] P. Barros, H. Ferreira, and M. Lemos, (2023), "Industrial organization of the maritime seismic acquisition market", gnpublication, *International Journal For Research In Business, Management And Accounting*, vol 9.

[5] Varian, H. R., Farrell, J. and Shapiro, C., (2004), "The Economics of Information Technology: An Introduction.", https://doi.org/10.1017/CBO9780511754166 *Cambridge University Press*.

[6] Shapiro, C., Varian, H. R., (1999), "The Economics of Information Technology: An Introduction.", https://doi.org/10.1017/CBO9780511754166 *Cambridge University Press*.

[7] Moody, D. and Walsh, P. (1999), "Measuring the Value Of Information-An Asset Valuation Approach.", ECIS, *Inproceedings*, pp. 496–512.

[8] Bakos, Y., Brynjolfsson, E., and Lichtman, D., (1999), "Shared information goods.", The University of Chicago Press, *The Journal of Law and Economics*, vol 42, n 1, pp. 117–156.

[9] Geng, X., Stinchcombe, M., and Whinston, A., (2005), "Bundling information goods of decreasing value.", INFORMS, *Management Science*, vol 51, n 4, pp. 662–667.

[10] Zhao, Z., *et al.*, (2011), "Geological exploration theory for large oil and gas provinces and its significance.", Elsevier, *Petroleum Exploration and Development*, vol 38, n 5, pp. 513–522.

[11] Damodaran, A., (2012), "Investment valuation: Tools and techniques for determining the value of any asset. *John Wiley & Sons*, vol 666.

[12] Andersen, H., (2007), "Value of information: reliability of 3D reflection seismology in exploration.", University of Stavanger, Norway, vol 6, n 1, pp. 30–41.

[13] Dias, M., (2004), "Valuation of exploration and production assets: an overview of real options models.", Elsevier, *Journal of petroleum science and engineering*, vol 44, n 1-2, pp. 93–114.

[14] Jafarizadeh, B. and Bratvold, R., (2015), "Oil and gas exploration valuation and the value of waiting.", Taylor & Francis, *The Engineering Economist*, vol 60, n 4, pp. 245–262.

[15] Shafiee, M., Alkali, B. and Baglee, D., (2019), "Decision support methods and applications in the upstream oil and gas sector. ", Elsevier, *Journal of Petroleum Science and Engineering*, vol 173, pp. 1173–1186.

[16] K. Ben-Naceur,(2019), "How the oil and gas industry is contributing to sustainability", SPE, *Journal of Petroleum Technology*, vol 71, pp. 38–39.

[17] T. Ahmad, and D. Zhang,(2020), "A critical review of comparative global historical energy consumption and future demand: The story told so far", Elsevier, *Energy Reports*, vol 6, pp. 1973–1991.

[18] M. Bashir, (2022), "Oil price shocks, stock market returns, and volatility spillovers: a bibliometric analysis and its implications", Springer, *Environmental Science and Pollution Research*, vol 29, pp. 22809–22828.

[19] D. Phan, *et al.*, (2022), "Crude oil price uncertainty and corporate investment: New global evidence", Elsevier, *Energy Economics*, vol 77, pp. 54–65.

[20] D. Phan, S. Sharma, and P. Narayan,(2015), "Oil price and stock returns of consumers and producers of crude oil", Elsevier, *Journal of International Financial Markets, Institutions and Money*, vol 34, pp. 245-262.

[21] M. Ilyas, *et al.*, (2021), "Economic policy uncertainty, oil price shocks and corporate investment: Evidence from the oil industry", Elsevier, *Energy*, vol 97, pp. 105–193.

[22] M. Berntsen, *et al.*, (2018), "Determinants of oil and gas investments on the Norwegian Continental Shelf", Elsevier, *Energy*, vol 148, pp. 904–914.

[23] D. Phan, S. Sharma, and P. Narayan,(2011), "Analysis and forecasts of investment scale and structure in upstream sector for oil companies based on system dynamics", Springer, *Petroleum Science*, vol 8, pp. 120–126.

[24] B. Zhang, and Q. Wang (2011), "Analysis and forecasts of investment scale and structure in the upstream sector for oil companies based on system dynamics", China University of Petroleum Beijing, *Petroleum Science*, vol 8, issue 1, pp. 120–126.

[25] S. Arora, (2020), "Investment Decision Making in the Upstream Oil Industry: An Analysis", http://ssrn.com/abstract=1983123, *Available at SSRN 1983123*.

[26] K. Lehn, P. Zhu, and K. Graduate, (2014), "Debt, Investment and Production in the U.S. Oil Industry: An Analysis of the 2014 Oil Price Shock", https://ssrn.com/abstract=2817123, *Available at SSRN 2817123*.

[27] T. Ahmad, and D. Zhang, (2020), "A critical review of comparative global historical energy consumption and future demand: The story told so far", Elsevier Ltd, *Energy Reports*, vol 6, pp. 1973-1991.

[28] R. Sutton, and A. Barto, (2018), "Reinforcement learning: an introduction", MIT press.

[29] D. Radovic, *et al.*, (2022), "Revealing Robust Oil and Gas Company Macro-Strategies using Deep Multi-Agent Reinforcement Learning", http://arxiv.org/abs/2211.11043.

[30] D. Radovic and L. Kruitwagen and C. Witt, (2020), "Revealing the Oil Majors' Adaptive Capacity to the Energy Transition with Deep Multi-Agent Reinforcement Learning ", In Climate Change AI Workshop@ NeurIPS 2020, pp. 9.

[31] X. Yang, *et al.*, (2022), "A Review: Machine Learning for Combinatorial Optimization Problems in Energy Areas ", MDPI, *Algorithms*, vol 6, issue 6.

[32] J. An, A. Mikhaylov, and N. Moiseev, (2019), "Oil price predictors: Machine learning approach", Econjournals, *International Journal of Energy Economics and Policy*, vol 9, pp. 1-6.

[33] V. Nanduri, and T. Das, (2007), "A reinforcement learning model to assess market power under auction-based energy pricing ", http://ssrn.com/abstract=1983123, *IEEE Transactions on Power Systems*, vol 22, issue 1, pp. 85-95.

[34] M. Rodríguez, *et al.*, "Bidders recommender for public procurement auctions using machine learning: Data analysis, algorithm, and case study with tenders from Spain ", Hindawi Limited, *Complexity*, vol: 2020, issue 3.

[35] F. Gu, Z. Jiang, and J. Su, (2021), "Application of Features and Neural Network to Enhance the Performance of Deep Reinforcement Learning in Portfolio Management ", In 2021 IEEE 6th International Conference on Big Data Analytics (ICBDA), *IEEE*, pp. 92–97.

[36] M. Janabi, (2022), "Optimization algorithms and investment portfolio analytics with machine learning techniques under time-varying liquidity constraints ", Emerald Group Holdings Ltd., *Journal of Modelling in Management*, vol 17, issue 3, pp. 864-895.

[37] H. Ghoddusi, G. Creamer and N. Rafizadeh, (2019), "Machine Learning in Energy Economics and Finance: A Review, "Elsevier, *Energy Economics*, vol 81, pp. 709-727.

[38] N. Vo, *et al.*, (2019), "Deep Learning for Decision Making and the Optimization of Socially Responsible Investments and Portfolio, "Elsevier, *Decision Support Systems*, vol 124, pp. 113097.

[39] Yoon, J., Arik, S. and Pfister, T., (2020), "Data valuation using reinforcement learning."International Conference on Machine Learning, *PMLR*, pp. 10842–10851

[40] Li, Y., (2017), “Deep reinforcement learning: An overview.”arXiv preprint, *arXiv:1701.07274*.

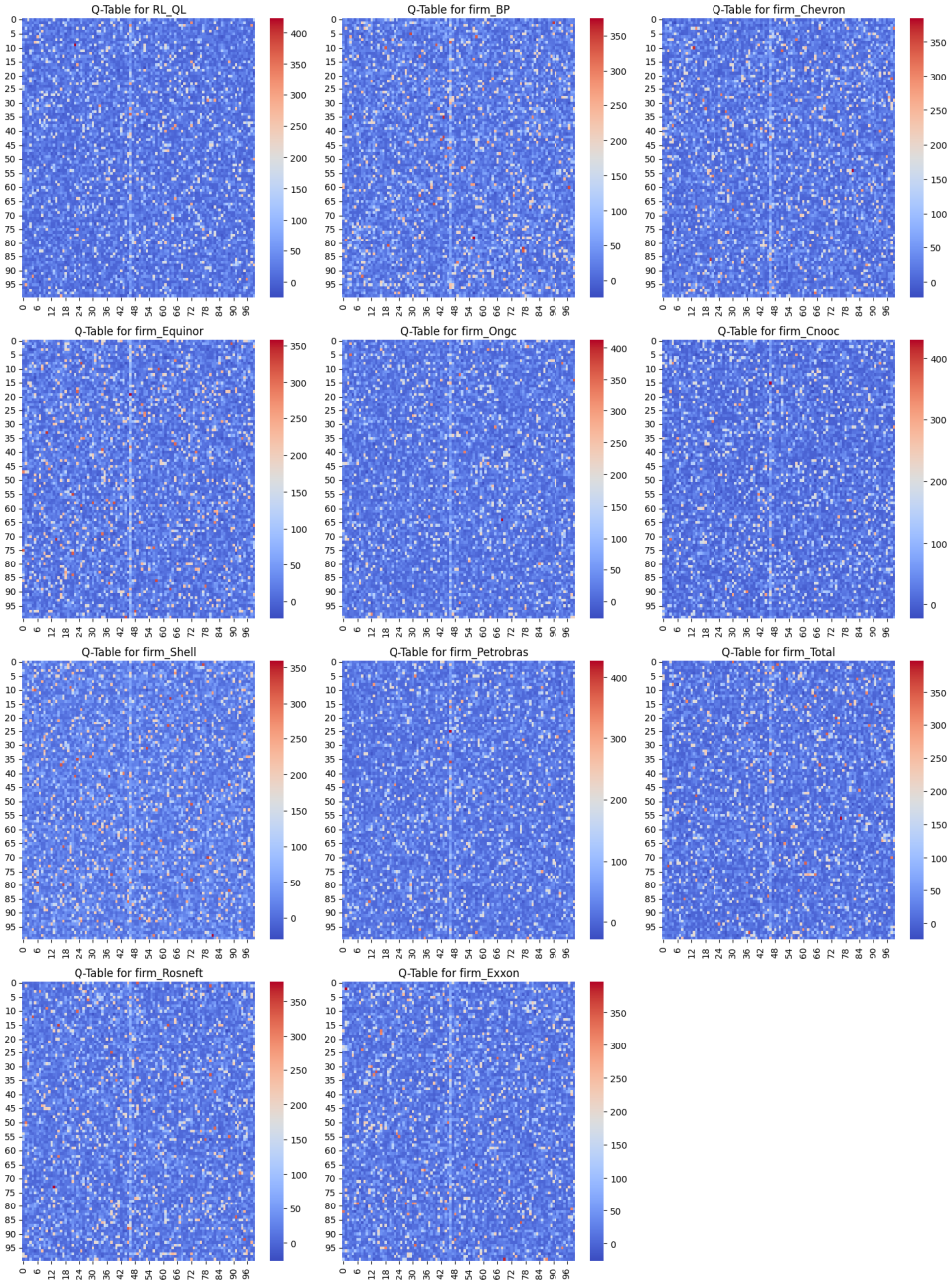
[41] Arulkumaran, K.,*et al.*, (2017), “Deep reinforcement learning: An overview.”IEEE Signal Processing Magazine, *IEEE*, vol 34, n 6, pp. 26–38.

[42] Mosavi, A., *et al.*, (2020), “Comprehensive review of deep reinforcement learning methods and applications in economics.”Mathematics, *MDPI*, vol 8, n 10, pp. 1640.

APPENDIX A

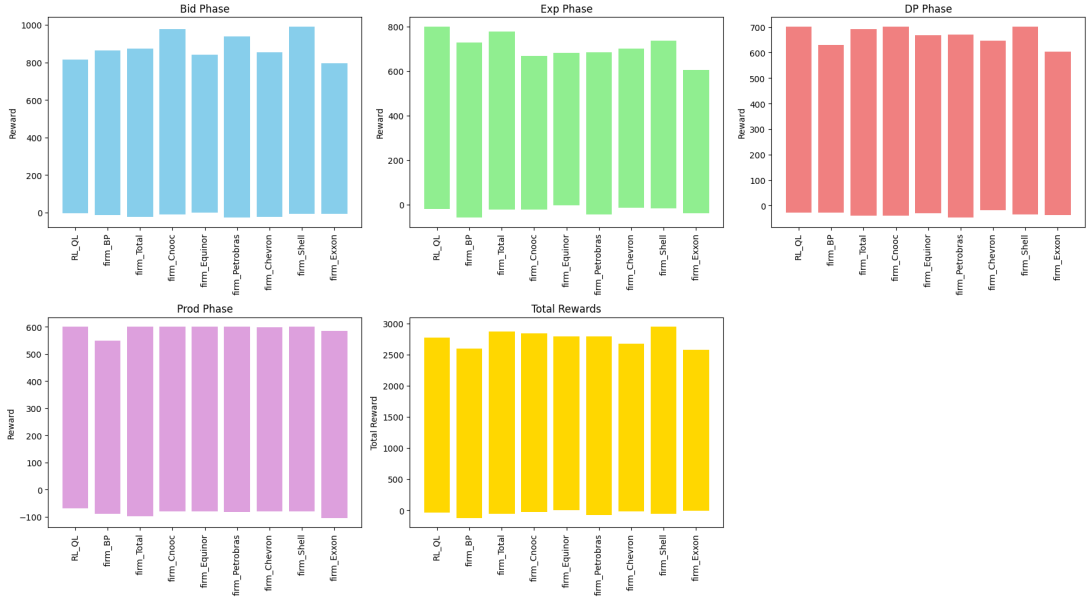
Supplementary Material

Figure 13: Q-Tables for All Firms $\epsilon = 0.5$ and 100k Rounds



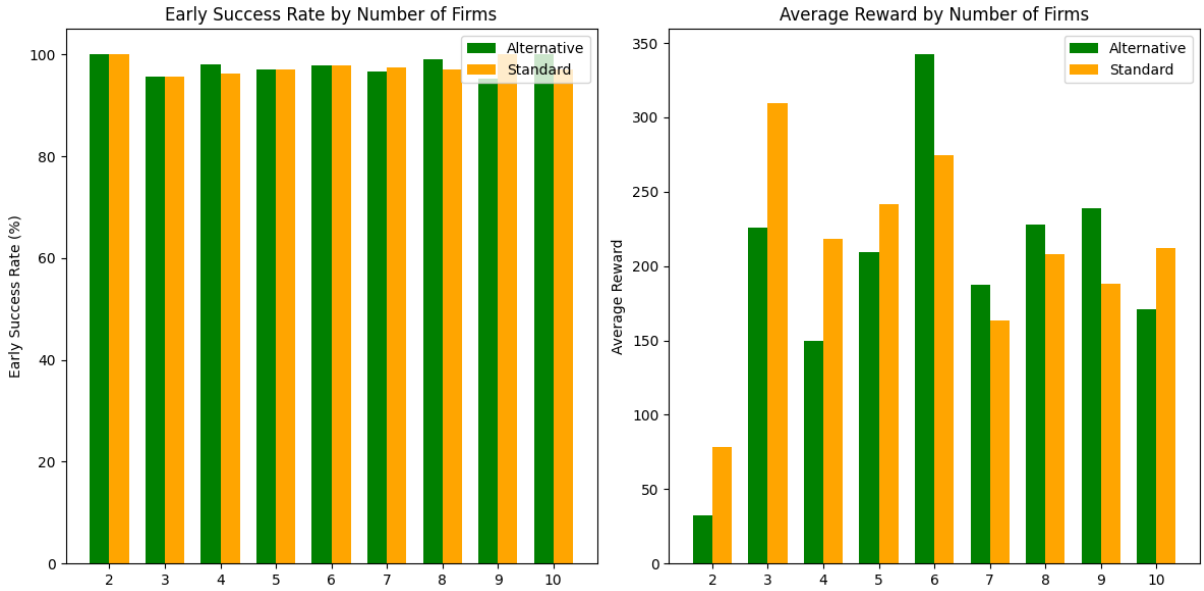
Note: Q-Tables for All Firms include the RL Agent trained by the model. For this value of e-greedy, we saw a mix between randomly state-action pairs and a line vertical pattern that represents the greedy solution found.

Figure 14: Rewards by Phase in one Competition Game



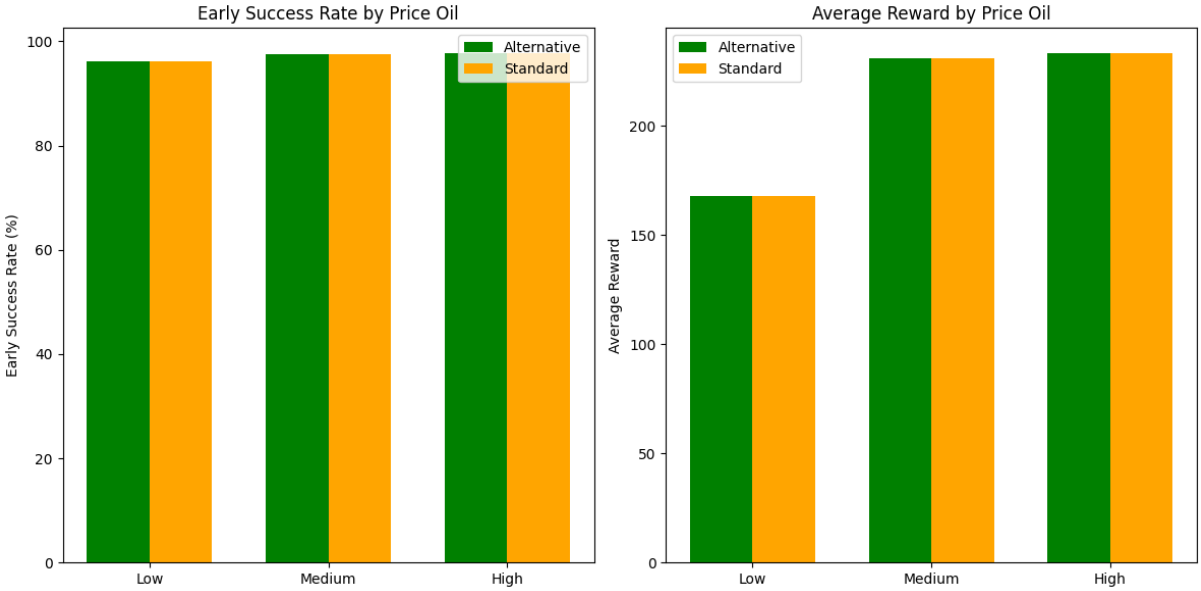
Note: We have the rewards by phase and the total values for one example of the Competition Game that defines the winner and the best rewards to be used in the update of the Q-Table.

Figure 15: Early Strategy Metrics Using DQN for Competition Feature



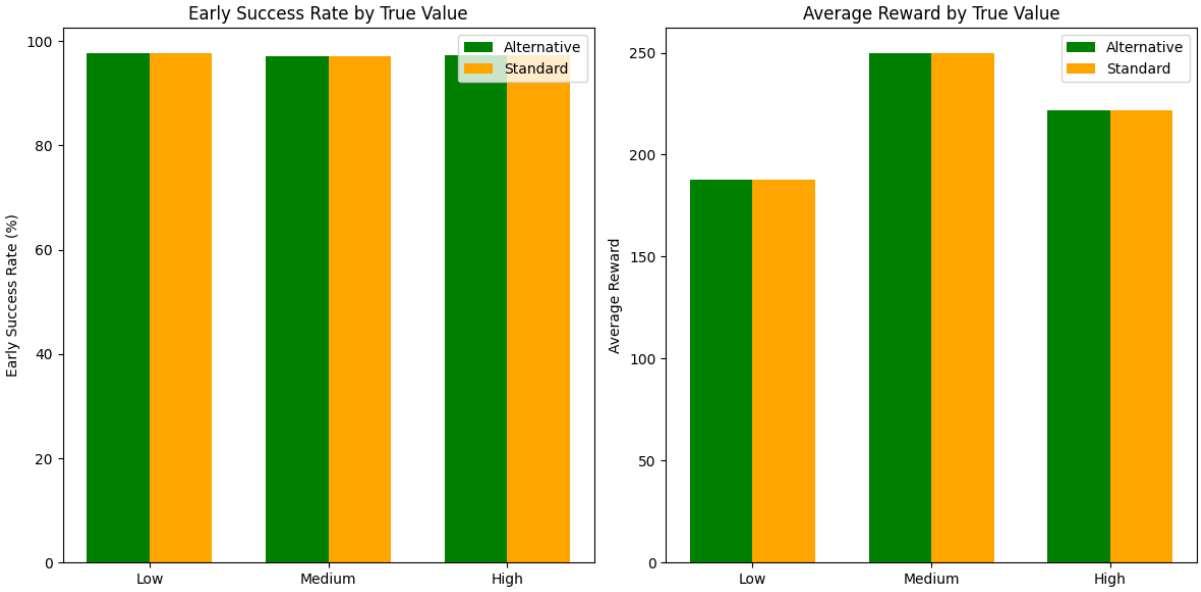
Note: On the left side, we have the Early Success Rate, and on the right side, the Average Reward segmented by Competition level. The number of firms varies between 2 and 10 during the bid phase competition.

Figure 16: Early Strategy Metrics Using DQN for Scenario Feature



Note: On the left side, we have the Early Success Rate, and on the right side, the Average Reward segmented by Scenario. The Low, Medium, and High in this context represent the heat, normal, or resilient market, which translates the level of demand, price, and volatility in the oil market.

Figure 17: Early Strategy Metrics Using DQN for Bid Feature



Note: On the left side, we have the Early Success Rate, and on the right side, the Average Reward segmented by Bid. In this context, the Low, Medium, and High represent the Bid with big, medium, and small leads in a log-normal scale.