Determining Optimal Incentive Policy for Decentralized Distributed Systems Using Reinforcement Learning

Elitsa Pankovska
elitsa.pankovska@gmail.com
University of Amsterdam

ABSTRACT

Cryptocurrencies have gained a lot of attention in recent years, mostly due to their decentralized manner of operation and their growth in value. However, a major drawback most of them possess is their high energy consumption. Current solutions to this problem have significant limitations: bringing back centralization and/or substituting the required energy with, e. g., storage space. This research aims to address the problem by investigating the use of a two-level deep reinforcement learning (RL) model to design incentive policies for green mining in cryptocurrencies. This is done by modeling one such system and creating an RL environment. Finally, by running simulations in it, we develop and test incentive policies, according to which cryptocurrency participants who use primarily renewable energy for their mining operations are more likely to add new blocks to the blockchain. Our results show that even when the green score of each crypto miner (determined by their energy mix) has relatively small importance (up to 0.3) in their selection probability, miners still shift towards green mining in order to increase their chance of being picked to validate cryptocurrency transactions and receive the corresponding rewards.

KEYWORDS

Reinforcement learning, Policy development, Blockchain, Renewable energy

1 INTRODUCTION

Decentralized systems such as blockchain-based cryptocurrencies have seen significant interest from both industry and academia over the last few years [18]. This interest is primarily accredited to the provision of conducting peer-to-peer transactions without the need for a trusted third party such as a bank or a financial institute.

Removing a trusted validator (bank or a financial institute) is a nontrivial problem due to the presence of attacks such as double-spending. In a double-spending attack, a user sends the same units of (digital) coins twice to users [1]. Without a trusted entity (for example, a bank), there is no easy way to validate if the sender did indeed have enough balance. To achieve this, blockchain and other similar decentralized distributed systems use incentive engineering to shape participant behavior: reward users who stop double-spend attacks and penalize the ones that support or prompt attacks on the network [2]. The most prominently used incentive engineering approach in blockchain systems is Proof of Work (PoW) which asks participants to spend computational cycles.

PoW is a form of adding new blocks of transactions to a cryptocurrency's blockchain [29]. The work, in this case, is generating a hash (a long string of characters) that matches the target hash for the current block. The crypto miner who does this wins the right to add that block to the blockchain and receive rewards.

PoW, however, has been widely criticized for its high electricity consumption. According to some recent reports, Bitcoin's network consumes a yearly amount of electricity comparable to that of the Republic of Ireland [20]. Current alternatives to PoW, such as Proof of Stake, solve this problem by coming up with ways that do not require as much or any computing power for the validation of transactions. Such approaches, however, are criticized for bringing back centralization and/or being less secure [17, 23].

The problem of finding PoW alternatives can benefit from more incentive engineering that focuses on or prompts green energy as a source. Reinforcement learning (RL) has the ability to assist in such situations where empirical or real world studies are not feasible or enough. RL has been used in recent research for policy design in the area of economics and finance and proves to overcome the limitations of current approaches [34]. For example, theoretical approaches are unable to capture the complexity of the real world, while empirical studies do not take into consideration the behavioral responses to policy behavior [34].

An RL model deals with these issues due to its possibility to run a large number of simulations, during which both the system and the agents adapt to the changing conditions to achieve optimal results. The agents take decisions based on rewards and punishments and learn strategies by interacting with the changing environment [27]. The system acts as a social planner that adapts to agent behaviors and seeks to achieve a particular policy objective [34].

This project aims to determine optimal incentive policy using a deep reinforcement learning-based simulation model to encourage the use of green energy for mining. An incentive policy can be any system, based on rewards and penalties, adopted to motivate user behavior. Additionally, we define as "green" all energy that is collected from renewable resources that are naturally replenished on a human timescale (such as wind or solar energy). In our case, the agents would be the so-called "miners" with the goal of earning Filecoins upon successfully mining a block, while the social planner would be the Filecoin system itself that strives towards green energy adoption while maintaining the system's reliability and security.

1.1 Research Question

We structure our research around this key research question:

Can a two-level deep reinforcement learning model be used to design incentive policies for green mining in cryptocurrencies with high energy consumption, such as Filecoin? The two levels in question are the following:

- Agents that attempt to maximise their reward.
- Social planner that attempts to increase the adoption of green energy use for mining through changes in incentive policies.

This research question can be further split into two sub-questions:

(1) How to model the system of Filecoin with a focus on its energy consumption?
(2) How to simulate the incentive policy of Filecoin using reinforcement learning techniques?

This paper is broken into 6 sections. We present a selection of background information and related work in Section 2. In Section 3, we provide a brief overview and decompose the challenge into individual steps, phases, and components and describe the background of the used technologies and how they were adapted to our specific use case. This section also contains information about the experimental setup and the used data sources. The results follow in Section 4. Some points of discussion and limitations of our study are provided in Section 5. In the end, we conclude our work and propose suggestions for future work in Section 6.

2 RELATED WORK

This section provides a brief overview of how blockchain technologies operate in 2.1 and introduces the Proof of Work concept, including alternatives thereof and their limitations in 2.2. In 2.3 we discuss the limitations of current approaches to modeling energy consumption in blockchains and provide our reasoning for choosing to work with the Filecoin system. Related work in the area of policy development using reinforcement learning is presented in 2.4. Finally, RL research in the blockchain domain is summarized in 2.5.

2.1 Blockchain Background

A blockchain is a distributed database that is shared among the nodes of a computer network. It consists of individual blocks, which are linked together. When a new block is created and filled, it is linked to the previously filled block (it includes that block's hash value) and can no longer be modified, forming a chain of data - the blockchain [6]. The main benefit of blockchain systems is that they provide security, anonymity, and data integrity without the need for a third-party organization in control of the transactions [32].

2.2 Proof of Work and Known Alternatives

To ensure that new blocks are added to the blockchain correctly, cryptocurrencies make use of consensus algorithms, such as Proof of Work (PoW)[26]. In the context of PoW, newly added transactions are sent to the memory pool[1] and nodes in the network compete between each other for getting the chance to add a new block to the blockchain. Generating a new block is done by finding a nonce value, which peers then include in a block. Finding this value is a computationally expensive operation - the first node in the network which manages to do this and mine the new block broadcasts the message to other nodes in the network who can verify its correctness [22].

Miners get rewarded upon generating a block thus incentivizing to support the network with computational power. The generated computational power of the network, also expressed in hash rate, protects the integrity of the PoW chain. Since verifying the correctness of a newly added block is done via a majority vote, the more hashing (computing) power in the network, the greater its security and its overall resistance to attack.

PoW solves the double-spend problem by guaranteeing that the transaction is not spent twice when it is included in a block. For such an attack to succeed, an actor would have to both alter previous blocks in their local copy and control simultaneously 51% or more of the copies of the blockchain, so that their version becomes the majority. This kind of attack, however, would cost so much computing power and money, that peers are far more economically incentivized in taking part in the network than attacking it.

PoW, however, has the following disadvantages [21]:[2]

- Its slow transaction speeds (10 minutes on average to build a new block)
- Mining often requires expensive equipment (due to the necessary computation power)
- High energy usage and carbon emissions

During this project, we focus on the high energy consumption of the system, although our proposed methodology also has the potential to solve the transaction speed problem. There is already a considerable amount of work done in this direction. Alternatives to PoW can be categorized into proof-based and voting-based algorithms [19]. The latter requires nodes to compare their results of the transaction validation, making the system less decentralized. For this reason, such algorithms are generally preferred less than proof-based approaches.

The most notable proof-based algorithm that lowers energy consumption during mining is Proof of Stake (PoS) [24]. The PoS model allows owners of the cryptocurrency to stake their coins, which gives them the right to add new blocks to the blockchain by validating transactions. However, if they validate a malicious transaction, they are penalized by losing some of their staked coins. The more coins a user stakes, the greater the chance of being chosen as a validator. This poses the problem that bigger stakeholders might have an excessive influence on the verification of transactions.

Alternative proof-based approaches attempt to replace energy with other types of resources. Dziembowski et al. [5] introduces the Proof of Space concept, which relies on data storage capacity. However, this alternative algorithm possesses the same problem as PoS - miners with an overwhelmingly large amount of storage have a much larger chance of validating transactions, making the system less decentralized. Also, this specific algorithm has one further flaw, namely, storage space is not used to store meaningful information. This in turn increases the demand for storage, the production of which also has a negative impact on the environment [12].

In contrast, we do not try to lower the energy consumption during the mining process per se, but rather demonstrate that tweaks in incentive policy can tilt participants in blockchain systems towards green mining solutions.

2.3 Energy Estimation in Cryptocurrencies

Simulating a cryptocurrency system with a focus on its energy consumption is a non-trivial task. A lot of research has gone into the estimation of the energy usage of PoW-based cryptocurrencies, especially that of Bitcoin [4, 13, 15]. Most of the research in this area delivers inconclusive/contradicting results, which are highly

[1]Memory pool is a cryptocurrency node's mechanism for storing information on unconfirmed transactions https://academy.binance.com/en/glossary/mempool

[2]This is a non-exhaustive list.

dependent on the current hash rate of the system in question[13]. The Bitcoin hash rate has increased significantly in recent years, increasing the mining difficulty and as a result also the energy consumption. Due to the difficulty in predicting how the hash rate of a cryptocurrency changes, energy estimation predictions for the future are also often inaccurate [15]. Also, the type of hardware used for mining plays a significant role in energy estimation and since mining hardware has developed a lot recently, the validity of research done just a couple of years ago is questionable [16].

For this reason, we base our work on the Filecoin cryptocurrency, the energy consumption of which does not depend on the hash rate of the system. While most cryptocurrencies' energy consumption stems from wasteful PoW computations, Filecoin uses an alternative mechanism called Proof-of-Spacetime. Filecoin is an open-source, public cryptocurrency and digital payment system intended to be a blockchain-based cooperative digital storage and data retrieval method. In the case of Filecoin, the miners can earn rewards by storing data for clients, and computing cryptographic proofs to verify storage across time [14], as well as by retrieving data.

Although this system performs useful computations, it still consumes a lot of electricity. The energy usage model Filecoin Green has come up with is tested by comparing estimated energy consumption to metered results for a limited number of the cryptocurrency participants. Moreover, lower and upper bounds are provided rather than just an average estimate. This makes their methodology more reliable than the energy estimation of PoW-based cryptocurrencies, which is why we focus on this blockchain system in particular.

2.4 Reinforcement Learning for Policy Development

Reinforcement learning (RL) encompasses a variety of machine learning algorithms that learn to make a sequence of decisions and form strategies. This is done by enabling an agent to maximize its cumulative rewards in an environment. Deep reinforcement learning (DRL) is a class of RL that uses neural networks to learn policies, which has seen a lot of interest in recent years [3], especially for the purpose of incentive engineering.

Zhan and Zhang [33] makes use of DRL to design the incentive mechanism for edge learning by capturing the tradeoff between latency and payment. Hu et al. [10] propose a reinforcement incentive learning method to develop incentive mechanisms to determine payment policies for crowdsourcing.

Zheng et al. [34] introduce a two-level DRL approach to learning dynamic tax policies, based on economic simulations in which both agents and a government learn and adapt. Although their research looks at equality and productivity in an economic context, it is very similar to our research problem in terms of number and types of agents and balancing different types of rewards.

Similarly to the authors, we focus on a two-level DRL - one where (instead of productivity and equality), we try to maximize productivity (mined filecoins) and the share of green energy sources used for mining. However, their economic simulation is an oversimplification of the real world, while we strive for a much more realistic representation of the Filecoin system. This is possible due to the smaller complexity of the system dynamics of a single blockchain system compared to those of the economy of an entire country. The key point in their research is that the social planners (which adapt the policies) and the normal agents have objectives that do not necessarily align. An example in the blockchain system would be when green energy is preferred by the social planner although it is more expensive than a traditional energy source.

In a subsequent study, Trott et al. [28] further prove the potential of AI in designing policies and its ability to deal with the complexity of the real world. They find that a two-level DRL model outperforms alternative approaches by taking into account both strategic planners and agents and interactions between them. For this reason, we believe that a similar approach can be utilized in designing incentive policies that can stimulate green mining solutions.

2.5 Reinforcement Learning in a Blockchain Context

To our knowledge, there is no existing research that utilizes RL in solving the problem of high energy consumption in PoW-based cryptocurrencies. Nevertheless, DRL has been used by Hou et al. [9] to study attacks on blockchain incentive mechanisms. The authors emphasize the vulnerability of such incentive mechanisms and prove that such can be exploited by rational users.

Yang et al. [31] use reinforcement learning to make the routing decisions of nodes in a wireless sensor network (WSN) and use blockchain to maintain the security of the system. They use an existing Proof of Authority algorithm, meaning that RL techniques are not used to increase the sustainability of the system, but rather to improve its trustworthiness. Similarly, Xiao et al. [30] choose blockchain to deal with selfish edge attacks and faked service record attacks in mobile edge computing (MEC). The authors propose a reinforcement learning (RL) based central processing unit (CPU) allocation algorithm for edge devices to allocate CPU resources for mobile devices in MEC, which results in lower energy consumption compared with a benchmark MEC scheme.

Jameel et al. [11] review RL techniques as a solution to some of the challenges of IIoT-blockchain networks, for example, minimizing forking events and improving the transactional throughput, which may result in a lower energy use. However, as the authors discuss, utilizing RL algorithms in real-time in blockchain systems (IIoT devices in their case) is an additional energy consumer.

None of the discussed papers uses RL for policy development of an alternative mechanism to PoW. Moreover, none focuses on the energy estimation of blockchain networks and especially in incentivizing green mining, which is the main focus of this study. Our goal is to design incentive policies for green mining in cryptocurrencies using RL rather than running RL algorithms in real-time on multiple devices in the network, which may improve system reliability [3] but is a further source of energy consumption.

3 METHODOLOGY

This section describes the methodology used to answer the proposed research questions. There are two core aspects of this research: modeling our blockchain system of choice and simulating its incentive policy using reinforcement learning. As already mentioned, most current alternatives to PoW bring back centralization

[3] as multiple studies suggest [11, 30, 31], however, none of them bases their work on cryptocurrencies.

and some of them even introduce a new type of waste (storage space in the case of Proof of Space). We use RL techniques for this study for three main reasons:

- We are able to implement an environment, which is a simplified representation of the real-world system, and monitor the behaviour of agents in the environment with respect to changing conditions.
- By using a two-level DRL model, where a social planner changes the miner selection policy, we can find a balance between two factors: the adoption of green energy in the system and the system reliability.
- In an RL simulation, agents can find ways to cheat the system, which we might not have anticipated initially, so we can take them into account to make improvements before introducing our solution to the real-world system.

3.1 System Modeling

An essential part of this research project is modeling the blockchain system - in our case, this would be Filecoin. The Filecoin ecosystem has in place many provisions that can help actors track the use of renewable energy such as the reputation system.[4] In our simulation, our goal is to establish if the inclusion of such green energy consideration in the calculation of the reputation system scores helps miners move to green sources of energy.

In order to do so, we need to model the Filecoin system while focusing primarily on the energy consumption aspect of it. Firstly, we make use of Filecoin Green's methodology of estimating energy consumption.[5] According to the authors, the main sources of energy consumption are sealing and storing data, which means that data retrieval operations are not a significant factor. Therefore, we only model the data storage in our system. The derived formula looks as follows:

$$ElectricalPower = (A \cdot (SealingRate) + B \cdot (RawCapacity)) \cdot PUE$$

Where:

- Electrical Power (watts) is the total electricity use
- A (Wh/byte) is the energy required to seal a sector
- Sealing Rate (bytes/hour) is the rate at which new data is being sealed, determined from on-chain proofs
- B (W/byte) is the electrical power required to store data over time
- Raw Capacity (bytes) is the amount of data stored, determined from on-chain proofs
- PUE (dimensionless) is the Power Usage Effectiveness, which is the ratio of total electrical power consumed to that consumed by useful IT processes such as sealing and storage

This model of energy consumption includes the three dominant processes which consume energy on Filecoin:

- Sealing, a one-time setup process onboarding data to the network
- Storage of files over time

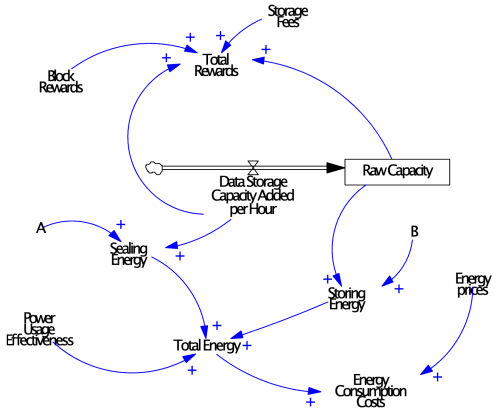

Figure 1: Filecoin System Modelling

- Overhead energy used to perform functions such as cooling and power conversion accounted for by Power Usage Effectiveness (PUE)

A (sealing energy), B (storage energy), and PUE are constants, for which Filecoin have come up with lower and upper bounds, as well as an average estimate. In our implementation, we use the upper bound values and therefore end up with electricity bills, which can also serve as an upper bound. An example that illustrates our reasoning is the following: in a system with lower energy costs, miners might be more willing to invest in green mining and still get a decent profit, whereas, in a scenario with higher bills, their profit might vanish if they also pay for renewable energy. Thus, if we get the expected results from our simulations, they would also be valid (if not better) in a scenario with lower energy costs. An overview of the system (created with the Vensim tool [6]) can be seen in Figure 1. By modeling the system using Vensim, we make sure that the units in the system match and our simulations run as expected. The plus signs in the figure indicate a positive correlation between the different variables.[7] Additionally, we run a four-month-long simulation and make a comparison between the simulated data and the real data as provided in the Filecoin energy dashboard.

Once we can estimate the data stored and the total energy used by a certain miner in the system, we can calculate both their rewards and their expenses. Block rewards are granted upon successfully sealing a new block of data and amount to 183.5$, which is equivalent to 21.3871FIL at the time of implementation.[8] Storage fees are calculated per byte per hour and amount to 3.057e-19$, again calculated using the current exchange rate between US$ and FIL.[9]

We gather data about the average data storage capacity added per hour from the Filecoin Energy Dashboard[10] and scale it down since our simulations are a lot smaller than the actual system.

A crucial feature of the system is the electricity prices, calculated per kWh[11] since they directly influence the profitability of the

[4]The Renewable Energy Purchased column is not shown by default, needs to be selected manually https://filrep.io/
[5]https://filecoin.energy/methodology
[6]https://vensim.com/
[7]For example, higher storage fees result in higher total rewards.
[8]Information about block rewards in real-time fetched from https://filfox.info/en
[9]Average storage fees calculation available here: https://file.app/
[10]https://filecoin.energy/
[11]publicly available at https://www.globalpetrolprices.com/electricity_prices/

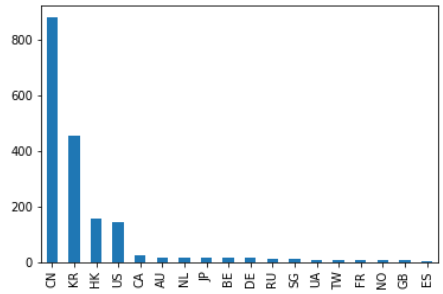

Figure 2: Geographic Distribution of Filecoin Miners

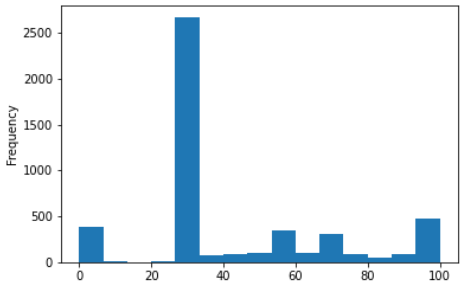

Figure 3: Distribution of Filecoin Miners' Reputation Scores

mining operations. They depend on the location where the miners operate. For this reason, we fetch all miners from the Filecoin reputation system (which provides information about their location) and calculate their geographic distribution (see Figure 2[12]). As we can see from the barplot, most miners are located in China, followed by South Korea, Hong Kong, and the United States.

Prices differ significantly, depending on the miner's location, e. g., the highest price is in Denmark (0.344$/kWh) and the lowest one is in Iran (0.005$/kWh). Some key statistics of the electricity prices can be seen in Table 1.

A miner with an electrical power of, e. g., 1kW consumes 1kWh of electricity per hour. Thus, the energy consumption costs (or electricity bills) are calculated on an hourly basis, by multiplying the electrical power in kW with the energy price at the respective location in $/kWh.

Table 1: Descriptive Statistics of the Energy Price per kWh in Different Countries

Summary: Energy Price per kWh in US$	
Mean	0.162154
Std	0.081045
Min	0.005000
25%	0.094000
50%	0.163000
75%	0.211500
Max	0.344000

3.2 Reinforcement Learning Implementation

In order to simulate the system, we use reinforcement learning and create a custom environment. We extend the implementation provided by the AI Economist[13], as it covers a similar use case: a two-level network, with a system that adjusts the policies, and agents that have the goal of maximizing their reward. We extend the framework by building a custom scenario, which includes how agents in the system are defined, what happens at each timestep of the simulations, and how the rewards are calculated. Additionally, we define the possible actions for the two types of agents: miners and a social planner (system), and how they influence the system dynamics. An overview of the environment can be seen in Figure 4.

Each miner in the system has two key features: a green score and a reliability score. They form a total score, according to which miners are picked to add blocks to the blockchain and earn rewards. The system plays the role of a planner, who can modify how the total score is calculated.[14] It can increase the importance of the green score or mostly trust the reliability of a miner, which is a static value for each miner. The goal of the system is to find a balance between the two to increase the share of green energy used for mining while still keeping the trust in the system (by picking reliable miners). The miners have the option to buy renewable energy certificates (RECs) to increase their green score. A REC is a market-based instrument that certifies the ownership of one megawatt-hour (MWh) of electricity generated from renewable energy sources.[15] We use RECs as a crude proxy for the use of renewable energy.

The reliability score is necessary because, without it, the system planner would simply set the green score importance to 1.0, in which case it is only natural that the agents would buy RECs. However, our purpose is to balance out the two aspects and find out what importance the green score should have as a minimum in order to motivate miners to use renewable energy. This score (also called reputation score in the Filecoin reputation system) is determined by 3 sub-scores: online reachability, committed sector proofs, and storage deals. Since how these scores are calculated is not of major importance to our study, we fetch the reliability scores from the Filecoin reputation system, calculate their distribution and assign static reliability scores to the miners in our environment with a matching distribution of scores (see Figure 3). We observe a large number of miners with a score of 30. Most of the other miners have a score of either 0, 50-60, or the maximum of 100.

Upon initializing the environment, the initial state of each miner has to be set. Firstly, we determine the location of each miner, which is chosen in a probabilistic way, where the probabilities are calculated using the real geographic distribution of nodes in the system (e. g., the probability of drawing China as a location is 0.4763).

[12] Countries with less than 5 miners are not represented in the plot, but are taken into account in the implementation.
[13] https://github.com/salesforce/ai-economist

[14] An example formula could be TotalScore = 0.9 * ReliabilityScore + 0.1 * GreenScore
[15] https://www.investopedia.com/terms/r/rec.asp

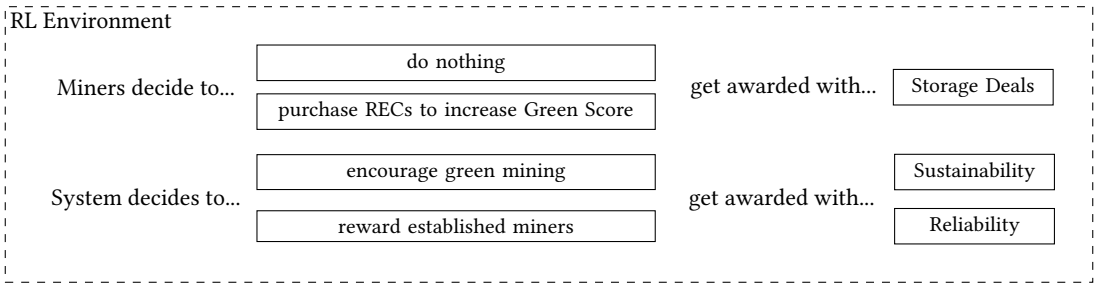

Figure 4: Reinforcement Learning Environment

This location determines the electricity price and the initial green score of the miner, which is equivalent to the share of electricity coming from renewables in the respective location [7]. As already mentioned, the reliability score is also drawn probabilistically. The initial total score is equal to the reliability, meaning that the initial importance of the green score is 0.0.

We also keep track of the energy consumed by the miner from the previous 24 timesteps (each timestep amounts to 1 hour), in the beginning of the simulation, it is set to an array of zeros.

The simulation works as follows: at each timestep, 10% of the agents (miners) are picked to add new blocks to the blockchain[16] (i. e., store the client's data). Each block amounts to 1TiB of data. This decision is made based on the total score of each miner - ones with higher scores have a bigger chance of being picked. After this selection procedure, the energy consumption of each agent is calculated using the methodology outlined above.

Each agent has the option to purchase RECs, which amount between 0 and 100% (with a step of 5%[17]) of their energy consumption in the previous timestep. The price of a REC amounting to 1kWh is 0.025$.[18] If they do not purchase any RECs, their green score remains the same. If they do, it goes up. Buying RECs increases costs, but also increases the probability of being picked to store new data and earn rewards as a result.

The green score is a weighted average of the miner's green scores from the previous 24 timesteps, where the weights correspond to the energy consumed in the respective timestep. The green score for each miner always ranges from their initial green score to 1.0. For example, since the share of electricity coming from renewables in Norway is already at more than 0.99 [7], a Norwegian miner cannot get a score of more than 1.0, even if they purchase 100% worth of RECs.

Every 24 timesteps (once a day), the system can adjust how the total scores are calculated. The green score importance can range between 0.0 and 1.0 with a step of 0.05.

At each timestep, the rewards for both the miners and the system are calculated. The miners' rewards are calculated by subtracting their electricity costs and the price of their RECs purchases from their total rewards: the block rewards and the storage fees. The system reward is an average of the system reliability and the system green score at the respective timestep. Let's assume that only one miner was picked at that timestep, which has a green score of 0.5 and a reliability of 0.7. This means that the system reward would just be the average of the reliability and the green score of that single miner - 0.6.

In this way, we can investigate what the importance of the green scores should be to incentivize green energy mining while keeping the trust in the system. Furthermore, the results will most likely depend on the location of the miners and their initial green scores: in some countries, electricity prices are lower so agents can have more available resources to purchase RECs.

We carry out the multi-agent training using the RLlib Proximal Policy Optimization algorithm which performs comparably or better than state-of-the-art approaches while being much simpler to implement and tune [25]. We use the default RLlib training parameters with a few small alterations:

- num_worker: 2
- num_envs_per_workers: 2
- train_batch_size: 4000
- sgd_minibatch_size: 4000
- num_sgd_iter: 1
- lr: 0.00001

We make use of distributed RL, where roll-out and trainer workers operate in tandem. The roll-out workers repeatedly step through the environment to generate and collect roll-outs[19] in parallel, using the actions sampled from the policy models on the roll-out workers or provided by the trainer worker. Trainer workers gather the roll-out data (asynchronously) from the roll-out workers and optimize policies. By setting the number of workers and environments to 2, we end up with a total of 4 environment replicas used to gather rollouts. We set the batch sizes to 4000 to keep the iteration time small.

3.3 Experimental Setup

We run two types of experiments in our study - one to get deep insights into the behaviour of all the individual miners and a large-scale experiment, which delivers more realistic results and is closer to the real-world system.

[16] In an environment with less than 10 agents, one agent is picked

[17] This results in an action space of 21 possible actions which is granular enough to give the agents a wide variety of options without increasing the training time significantly.

[18] Data presented at the REMC 2021 shows a steep increase in prices of RECs - from 0.0015$ in Dec 2020 to 0.007$ in Aug 2021[8]. More recent price statistics are not available, but one of the largest providers Sterling Planet offer RECs in most US states for 0.015$/kWh with their most expensive REC being 0.03$/kWh, see http://sterlingplanet.com/regional-green-power-program/

[19] A roll-out is a single played out episode in the environment with a set number of timesteps

We set a seed of 0 for all random choice operations. Firstly, we train an RL environment with 4 miners and one planner, which runs for 100 timesteps (each timestep is equivalent to 1 hour), where miners can decide to buy RECs at each timestep, while the planner can adjust the green score importance only once a day (every 24 timesteps). We train the agents in our environment for 10 iterations.

This setup has been especially useful to realize weaknesses of our implementation and modify it accordingly. For example, the green score of a miner is calculated using the RECs they have bought on the previous day (24 timesteps). However, the REC purchases are calculated as a percentage of the energy consumption in the respective timestep. As a result, if each timestep has an equal importance in calculating the green score, miners would buy a lot of RECs in timesteps with low energy consumption and less/none when they consume a lot of energy. For this reason, we have modified the calculation of the miners' green scores by using the energy consumption in the respective timesteps as weights.

Secondly, we train an RL environment with 100 miners and one planner, which runs for 336 timesteps (2 weeks). Again, miners buy RECs each hour while the planner adjusts the policy once a day. The training is done for 20 iterations.

We evaluate our results by running one more episode of the simulation[20] with the already trained agents and observe the agent behaviour and the system metrics.

3.4 Evaluation Methods

The evaluation of our work is done in two steps. Firstly, our goal is to model the Filecoin system in a relatively simple way, while maintaining the key characteristics of the system and capturing the energy consumption as precisely as possible. We do this by simulating a four-month period using Vensim and comparing the simulated energy consumption with the datasets provided by Filecoin.

Secondly, we evaluate the results of our simulations (the created incentive policy) by comparing it to the current use of the Filecoin reputation system, which does not incorporate green energy purchases in the calculation of the miners' rankings. This would mean estimating how renewable the real system is compared to the system we end up with in our simulations. We would like to investigate to what extent miners would switch to green energy sources if they are not motivated to do so by the selection policy. Since our goal is to have a policy that balances system reliability and greenness, we run two additional experiments:

(1) A system with a green score importance of 0.0, where only the miner reliability forms the selection probability.
(2) A system with a green score importance of 1.0, where only the miner green score forms the selection probability.

We compare the results of these two experiments with our system in which both factors play a role in the miner selection policy using two evaluation metrics.

- System Green Score: a weighted average of the green scores of the individual miners, where the weights correspond to the amount of data each miner is storing.

[20]In a new episode, the environment is reset and the simulation is run for a certain number of timesteps: 100 in the small-scale experiment and 336 in the large-scale experiments

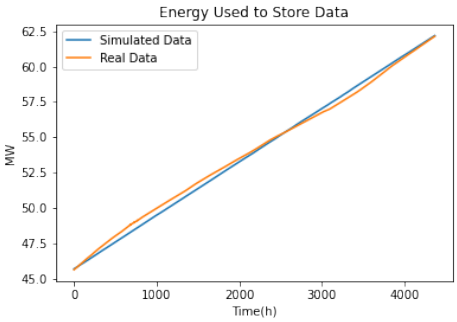

Figure 5: Comparison of the Data Storage of the Filecoin System and System Simulation

- System Reliability: a weighted average of the reliability scores of the individual miners, where the weights correspond to the amount of data each miner is storing.

4 RESULTS

In this section we present the results of our study, starting with the system modeling, a qualitative (from the small-scale experiments) and quantitative (from the large-scale experiments) evaluation.

4.1 System Modelling Validation

The model of the system we have created is based on the modeling done by Filecoin Green and adheres to the design outlined in their document.[21]

Furthermore, Figure 5 shows a comparison between the real data, fetched from the official Filecoin energy dashboard [22] and the artificial data we have simulated using Vensim. We compare the energy used to store data in particular and can see a very close fit. Since all other variables in the system model are statically set and come from real-world data (electricity prices, storage fees), we believe that the model of the system is accurate.

4.2 Qualitative Evaluation

In this subsection, we show the results of our small-scale experiments with just 4 miners, in which we are able to observe the behaviour of each miner. However, this setup is less suitable to derive conclusions for the real-world system, mainly because we can often end up in a special case, which will be discussed below.

The behaviour of the trained agents can be seen in Figures 7 and 8, which represent one episode of 100 timesteps. We can see that the agents have indeed learned that they need to buy RECs and in turn increase their green scores to get more storage deals. For example, at around timestep 30 Agent 2 comes to a halt - it barely gets to increase its total storage for a longer period of time. We can see that around timestep 60 the green score of Agent 2 starts to increase (which means the agent has decided to buy more RECs) and a few timesteps later their amount of data storage also increases. The importance of the green score at these timesteps is just 0.35. This suggests that the reliability of the system does

[21]https://github.com/redransil/filecoin-energy-estimation
[22]https://filecoin.energy/

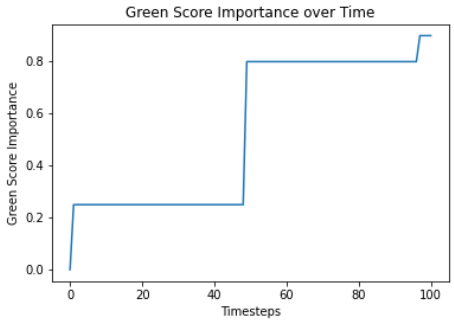

Figure 6: Green Score Importance over Time

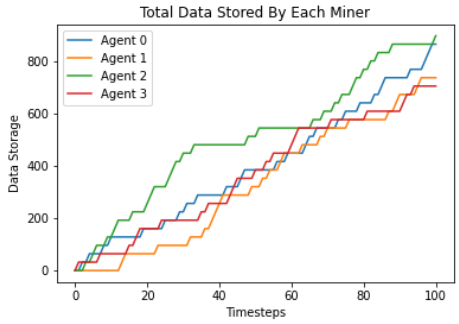

Figure 7: Total Data Storage of Each Miner over Time

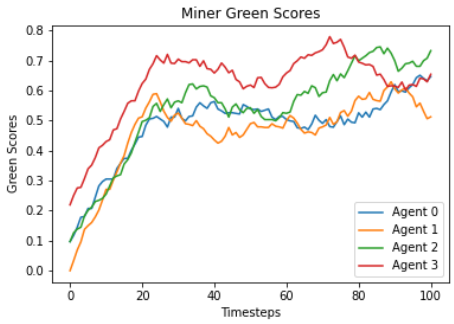

Figure 8: Green Score of Each Miner over Time

not need to be sacrificed in order to incentivize miners to adopt green energy mining. As a result, our research demonstrates that including similar green scores (even with a marginal importance) in the selection algorithms used by cryptocurrencies can motivate crypto miners to increase the percentage of renewable energy they use in their mining operations.

In this subsection, we also consider a special case of the small-scale system. Due to the small number of miners and the large probability of having a reliability score of 0.3 (see Figure 3), there is a large chance that all miners end up with the same reliability. In such a system, the green score importance determined by the planner increases to 0.9 to incentivize miners to buy as much RECs as possible and in this case increase system sustainability as much as possible, since system reliability doesn't change (see Figure 6). Also, miners' final green scores converge between 0.7 and 0.9, whereas in the generic case, they range between 0.5 and 0.75.

4.3 Quantitative Evaluation

Here, we present the results of a much larger experiment with a system of 100 miners, for which we keep track of the whole system metrics rather than those of individual system participants.

From Table 2 it can be seen that the dynamic green score importance, determined by the AI social planner, achieves the highest average percentage between the green and reliability scores. When the green score importance is statically set to 0.0, the agents cannot find a link between their REC purchases and their success in getting picked to store data. Therefore, miners generally purchase less RECs or purchase them mostly when their energy consumption is low (and in turn the price of the RECs is also low). In this scenario, miners are chosen only based on their reliability, so the system reliability score is as high as it can get (it serves as an upper bound). In the opposite case, only the green score is important so naturally, miners buy a lot more RECs and in turn, the system green score is as high as possible (it is not 100% because for most miners maintaining such a high score would result into higher costs than rewards). In the case when the social planner adjusts the green score importance, neither metric suffers dramatically and the balance between the two is kept. As seen in Figure 3, most miners have a reliability score of 30, so the relatively low average is understandable.

We have also run the same experiment for significantly more iterations (50), which resulted in a policy with a rather high green score importance, where the AI dynamic importance results get closer to the metrics from the system with a green score importance of 1.0. We believe the reason for this is that reliability of the system is capped at around 56%, due to these scores being static, so the planner sees more potential in increasing the average score by increasing the green score of the system as much as possible (which is theoretically capped at 100%).

5 DISCUSSION

In this section, we discuss potential answers to our research questions, provide points of discussion, and analyse the limitations of our work.

5.1 Research Outcome and Comparison to Existing Approaches

The system of Filecoin is modeled using the official energy consumption methodology, combined with data about the reward system of the cryptocurrency and real-world data about electricity prices in different locations. A key aspect of the model is taking into account the geographic distribution of nodes in the system since the location of a miner determines the electricity price and the share of electricity coming from renewables (which determines the miners' initial green scores).

We simulate the incentive policy of Filecoin using RL techniques by building an RL environment, in which participants in the system have two scores, a reliability and a green score, which determine their probability of getting picked to add new blocks to the

Table 2: Evaluation Results

Green Score Importance	System Green Score	System Reliability Score	Average Score
0.0	63.52%	**56.03%**	59.775%
1.0	**80.1%**	41.86%	60.98%
Dynamic	76.05%	47.75%	**61.9%**

blockchain. We introduce a social planner, who tries to balance the system reliability and the green score of the system by changing the importance of the latter one in the selection policy. We show that even with a relatively low green score importance miners are still motivated to increase the share of electricity coming from renewables while keeping the system's reliability.

Since there is no existing research on developing incentive mechanisms for mining using renewable energy sources, it is difficult to compare our results to other studies. For this reason, we came up with our own baselines to validate our work. As already mentioned, our solution surpasses the lower bounds in each of the scenarios with a static green score importance and achieves a higher average result than both of them. We believe that in a system where the reliability scores are not static (and in turn, the reliability of the whole system has a higher upper bound), an even lower green score importance can drive miners towards green mining.

In the current Filecoin reputation system, the top 70 highest-ranked miners have no record of renewable energy purchases.[23] Judging by our experiments, including these purchases as part of the reputation score and using it as a selection policy (where miners with higher scores have a larger chance of getting picked to store data) may motivate them towards more RECs purchases.

We can compare our solution to other alternatives to PoW in terms of decentralization and transaction speed. As already mentioned, Proof of Stake and Proof of Space have a common flaw - the more coins you stake or the more storage space you provide, respectively, the larger your chance of getting picked to validate transactions. In our solution, both the reliability score and the green score are bounded between 0.0 and 1.0, and theoretically, all miners can achieve a 1.0 in both scores[24], meaning that no single miner can have an overpowering influence over the system by having a much larger probability of being picked than others. Since our proposed selection algorithm does not require computationally expensive operations, the transaction speed can be increased significantly.

5.2 Limitations

We are aware that in a simulation, we can only capture a small number of the system dynamics. Therefore, our work comes with certain limitations.

Firstly, the observed behaviour of the miners in the RL simulation is highly dependent on the calculation of the rewards, which are formed by calculating the electricity costs, RECs purchases, storage fees, and block rewards - all of which are static variables, corresponding to the current situation. In our simulation, miners can make a profit even when purchasing a decent amount of RECs when taking into account the current price of RECs and the current value of Filecoin. These results may change in the future if the value of Filecoin drops significantly, the price of RECs soars, or even with a change in electricity costs. This means that while our results are valid for the current situation, this might change in the future. This limitation is closely linked to the overall profitability of crypto mining, which is difficult to predict for most if not all cryptocurrencies. While this problem is unavoidable, if one could predict the mentioned variables, the RL simulations can be run with minor changes to see how the situation might develop in the future.

Also, purchasing RECs is overall not seen as the best way of ensuring that the electricity a miner has consumed comes from renewables. Better alternatives are, e. g., producing green energy yourself by purchasing solar panels. This, however, would increase the complexity of our simulation significantly since how much energy such a setup produces depends on various factors, but mainly location and weather. Due to the scope of the study, a scenario in which a miner has the option to make a high upfront investment in hardware (solar panels) but then lower their electricity bills is left as a potential future work.

There are a couple of further limitations of this study: in our simulation, the storage space miners provide is theoretically infinite, which does not correspond to the real world. However, since we run the simulations for a limited number of timesteps, this has not proven problematic in our use case. Nevertheless, the simple solution would be to set a maximum amount of storage for each miner - once a miner has achieved it, they would no longer "play in the selection competition".

Also, in the Filecoin reputation system, the reliability scores are determined by various factors: online reachability, committed sector proofs, and storage deals. Since implementing the whole Proof-of-Spacetime mechanism in an RL environment would mean implementing the whole Filecoin system from scratch, it becomes difficult to calculate these scores in real-time during the simulations. Also, since we are not aware of the associated costs with improving these scores, we have decided to keep them static, using the distribution of scores from the real system. In future work, it is possible to also make this score non-static, by e. g., defining the cost of being online or giving miners the option to cheat by not storing data.[25]

6 CONCLUSION

To our knowledge, this is the first study that uses an RL-approach in incentive policy design in the cryptocurrency domain. Moreover, research done in the field usually has one main flaw - bringing

[23] Upon sorting miners by reputation score, the "Renewable Energy Purchased" column is N/A for the first 70 miners, see: https://filrep.io/

[24] The reliability score in the implementation is statically set using the probability distribution of scores in the real system, but theoretically, it could be 1.0 for all miners in the system.

[25] In such a case, they would not have to pay the electricity costs associated with sealing and storing data, but their reliability score would drop.

back centralization [23], which our solution does not possess. The proposed method incentivizes crypto miners to use renewable energy by introducing a green score, ranging from 0 to 1.0, which increases the probability of being picked to add new blocks to a blockchain. We prove that even when the green scores do not have a large importance in the selection policy, miners still tend to move towards green mining. Moreover, since the green score is capped at 1.0, we make sure that individual miners do not have such an overwhelmingly higher score than others that they have major control over the system (something that can happen in Proof of State or Proof of Space-based cryptocurrencies).

The validation results show that the trained policy using multi-agent RL techniques is able to balance system reliability and the green score of the system, achieving a higher average between the two than both the current selection policy (where green scores are not considered at all) and a static policy (where only the miners' green scores matter).

Despite the promising results, several limitations should be considered when applying the result in the real system. Firstly, purchasing RECs is neither the only nor the recommended way of proving renewable energy consumption. Moreover, the validity of the results is closely linked to the current value of Filecoin and the price of RECs and electricity. While the first limitation can be overcome by introducing other means of increasing one's green score (e. g., by producing renewable energy using purchased hardware) in the RL simulations, predicting/modeling the price of cryptocurrencies has proven to be a very complex task. However, the work we have done in this study can be easily adapted to changes in RECs and electricity prices and carried through again with minor modifications. Furthermore, the same approach can be used to modify the incentive policies of other cryptocurrencies. These three aspects will be covered in future work.

REFERENCES

[1] Nujud A. Alabdali, Mohammed Abdullatif Alzain, Mehedi Masud, Jehad F. Al-Amri, and Mohammed Baz. 2020. BITCOIN AND DOUBLE-SPENDING: HOW PAVING THE WAY FOR BETTERMENT LEADS TO EXPLOITATION.

[2] Joseph Bonneau, Andrew K. Miller, Jeremy Clark, Arvind Narayanan, Joshua A. Kroll, and Edward W. Felten. 2015. perspectives on Bitcoin and second-generation cryptocurrencies.

[3] Matthew Botvinick, Sam Ritter, Jane X. Wang, Zeb Kurth-Nelson, Charles Blundell, and Demis Hassabis. 2019. Reinforcement Learning, Fast and Slow. *Trends in Cognitive Sciences* 23, 5 (2019), 408–422. https://doi.org/10.1016/j.tics.2019.02.006

[4] Alex de Vries. 2018. Bitcoin's Growing Energy Problem. *Joule* 2, 5 (2018), 801–805. https://doi.org/10.1016/j.joule.2018.04.016

[5] Stefan Dziembowski, Sebastian Faust, Vladimir Kolmogorov, and Krzysztof Pietrzak. 2015. Proofs of Space. In *Advances in Cryptology – CRYPTO 2015*, Rosario Gennaro and Matthew Robshaw (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 585–605.

[6] Kurt Fanning and David P Centers. 2016. Blockchain and its coming impact on financial services. *Journal of Corporate Accounting & Finance* 27, 5 (2016), 53–57.

[7] Max Roser Hannah Ritchie and Pablo Rosado. 2020. Energy. *Our World in Data* (2020). https://ourworldindata.org/energy.

[8] Jenny Heeter, Eric O'Shaughnessy, and Rebecca Burd. 2021. Status and Trends in the Voluntary Market (2020 Data). (10 2021). https://www.osti.gov/biblio/1826295

[9] Charlie Hou, Mingxun Zhou, Yan Ji, Philip Daian, Florian Tramèr, Giulia C. Fanti, and Ari Juels. 2021. SquirRL: Automating Attack Analysis on Blockchain Incentive Mechanisms with Deep Reinforcement Learning. In *NDSS*.

[10] Zehong Hu, Yitao Liang, Jie Zhang, Zhao Li, and Yang Liu. 2018. Inference Aided Reinforcement Learning for Incentive Mechanism Design in Crowdsourcing. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/f2e43fa3400d826df4195a9ac70dca62-Paper.pdf

[11] Furqan Jameel, Uzair Javaid, Wali Ullah Khan, Muhammad Naveed Aman, Haris Pervaiz, and Riku Jäntti. 2020. Reinforcement Learning in Blockchain-Enabled IIoT Networks: A Survey of Recent Advances and Open Challenges. *Sustainability* 12, 12 (2020). https://doi.org/10.3390/su12125161

[12] Hongyue Jin, Kali Frost, Ines Sousa, Hamid Ghaderi, Alex Bevan, Miha Zakotnik, and Carol Handwerker. 2020. Life cycle assessment of emerging technologies on value recovery from hard disk drives. *Resources, Conservation and Recycling* 157 (2020), 104781.

[13] Sinan Küfeoğlu and Mahmut Özkuran. 2019. Bitcoin mining: A global review of energy and power demand. *Energy Research & Social Science* 58 (2019), 101273.

[14] Protocol Labs. 2017. Filecoin: A Decentralized Storage Network. (2017). https://filecoin.io/filecoin.pdf

[15] Jingming Li, Nianping Li, Jinqing Peng, Haijiao Cui, and Zhibin Wu. 2019. Energy consumption of cryptocurrency mining: A study of electricity consumption in mining cryptocurrencies. *Energy* 168 (2019), 160–168. https://doi.org/10.1016/j.energy.2018.11.046

[16] Eric Masanet, Arman Shehabi, Nuoa Lei, Harald Vranken, Jonathan Koomey, and Jens Malmodin. 2019. Implausible projections overestimate near-term Bitcoin CO2 emissions. *Nature Climate Change* 9, 9 (2019), 653–654.

[17] Mahdi H. Miraz, Peter S. Excell, and Khan Sobayel. 2021. Evaluation of Green Alternatives for Blockchain Proof-of-Work (PoW) Approach. *Annals of Emerging Technologies in Computing* (2021).

[18] Bhabendu Kumar Mohanta, Debasish Jena, Soumyashree S. Panda, and Srichandan Sobhanayak. 2019. Blockchain technology: A survey on applications and security privacy Challenges. *Internet of Things* 8 (2019), 100107. https://doi.org/10.1016/j.iot.2019.100107

[19] Giang-Truong Nguyen and Kyungbaek Kim. 2018. A Survey about Consensus Algorithms Used in Blockchain. *J. Inf. Process. Syst.* 14 (2018), 101–128.

[20] Karl J. O'Dwyer and David Malone. 2014. Bitcoin mining and its energy footprint. In *25th IET Irish Signals Systems Conference 2014 and 2014 China-Ireland International Conference on Information and Communications Technologies (ISSC 2014/CIICT 2014)*. 280–285. https://doi.org/10.1049/cp.2014.0699

[21] Zhiqiang Ouyang, Jie Shao, and Yifeng Zeng. 2021. PoW and PoS and Related Applications. *2021 International Conference on Electronic Information Engineering and Computer Science (EIECS)* (2021), 59–62.

[22] Amitai Porat, Avneesh Pratap, Parth Shah, and Vinit Adkar. 2017. Blockchain Consensus: An analysis of Proof-of-Work and its applications.

[23] Ashish Rajendra Sai, Jim Buckley, Brian Fitzgerald, and Andrew Le Gear. 2021. Taxonomy of centralization in public blockchain systems: A systematic literature review. *Information Processing & Management* 58, 4 (2021), 102584. https://doi.org/10.1016/j.ipm.2021.102584

[24] Fahad Saleh. 2020. Blockchain Without Waste: Proof-of-Stake. *Information Systems & Economics eJournal* (2020).

[25] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[26] S Sheikh, RM Azmathullah, and F Rizwan. 2018. Proof-of-work vs proof-of-stake: a comparative analysis and an approach to blockchain consensus mechanism. *International Journal for Research in Applied Science & Engineering Technology* 6, 12 (2018), 786–791.

[27] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA.

[28] Alexander Trott, Sunil Srinivasa, Douwe van der Wal, Sebastien Haneuse, and Stephan Zheng. 2021. Building a Foundation for Data-Driven, Interpretable, and Robust Policy Design using the AI Economist. *Machine Learning eJournal* (2021).

[29] Daniel van Flymen. 2020. *Proof of Work*. Apress, Berkeley, CA, 39–53. https://doi.org/10.1007/978-1-4842-5171-3_4

[30] Liang Xiao, Yuzhen Ding, Donghua Jiang, Jinhao Huang, Dongming Wang, Jie Li, and H. Vincent Poor. 2020. A Reinforcement Learning and Blockchain-Based Trust Mechanism for Edge Networks. *IEEE Transactions on Communications* 68, 9 (2020), 5460–5470. https://doi.org/10.1109/TCOMM.2020.2995371

[31] Jidian Yang, Shiwen He, Yang Xu, Linweiya Chen, and Ju Ren. 2019. A Trusted Routing Scheme Using Blockchain and Reinforcement Learning for Wireless Sensor Networks. *Sensors* 19, 4 (2019). https://doi.org/10.3390/s19040970

[32] Jesse Yli-Huumo, Deokyoon Ko, Sujin Choi, Sooyong Park, and Kari Smolander. 2016. Where is current research on blockchain technology?—a systematic review. *PloS one* 11, 10 (2016), e0163477.

[33] Yufeng Zhan and Jiang Zhang. 2020. An Incentive Mechanism Design for Efficient Edge Learning by Deep Reinforcement Learning Approach. In *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*. 2489–2498. https://doi.org/10.1109/INFOCOM41043.2020.9155268

[34] Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C. Parkes, and Richard Socher. 2021. The AI Economist: Optimal Economic Policy Design via Two-level Deep Reinforcement Learning. *ERN: Efficiency; Optimal Taxation (Topic)* (2021).