On Blockchain We Cooperate: An Evolutionary Game Approach

XINYU TIAN[†‡] and LUYAO ZHANG[*†‡§]

Blockchain is the trust machine in cyberspace that supports cooperation by consensus protocols. However, studies on consensus protocol in computer science ignore the incentives that could affect agent behaviors. An emerging literature in game theory introduces rational agents and solution concepts to study equilibrium outcomes of various consensus protocols. However, the existing studies with rational agents are limited in generalizability and are far from providing guidance for future designs of consensus protocols. We abstract a general Byzantine consensus protocol as a general game environment in extensive form, apply bounded rationality to model agent behaviors, and solve the initial conditions for three different stable equilibria. Our research contributes to literature across disciplines, including Byzantine consensus protocol in computer science, game theory in economics on blockchain consensus, evolutionary game theory at the intersection of biology and economics, and bounded rationality at the interplay between psychology and economics. Finally, our research guide future designs of consensus protocols to achieve desirable outcomes by evaluating incentives choices.

Keywords: cooperation, Byzantine fault tolerance, bounded rationality, evolutionary game theory, evolutionary stable strategy, blockchain consensus

We have no eternal allies, and we have no perpetual enemies. Our interests are eternal and perpetual, and those interests are our duty to follow.— Lord Palmerston, the mid-19th century British Prime Minister

[*]Corresponding authors: email: lz183@duke.edu, institutions: Data Science Research Center and Social Science Division, Duke Kunshan University.
[†]By the alphabetical order of the last name. Duke Kunshan University, No. 8 Duke Ave, Kunshan, Suzhou, Jiangsu, China, 215316.
[‡]Also with SciEcon CIC, London, United Kingdom, WC2H 9JQ
[§]ORCID: https://orcid.org/0000-0002-1183-2254

Authors' address: Xinyu Tian; Luyao Zhang.

1 INTRODUCTION

Cooperation is an evolutionary process that is fundamental for human prosperity [noa, 2018]. The mechanism of cooperation has been widely studied in a variety of disciplines such as biology [Koduri and Lo, 2021], psychology [Henrich and Muthukrishna, 2020], economics [Fehr and Fischbacher, 2004, Fehr et al., 2002, Fehr and Gächter, 2000], and computer science [Nisan, 2007]. Human evolution in the past one hundred years have witnessed marvellous advancements in Artificial Intelligence (AI) [Horvitz, 2016]. The advancement is integrating cyberspace (CS), physical space (PS), and social space (SS), into the Cyber-Physical-Social System (CPSS), which expands extraordinarily the territories of human civilizations [Wang et al., 2019]. However, the integration also brings new challenging issues for cooperation. Blockchain further enables cooperative AI [Dafoe et al., 2020] by implementing a consensus process. Since [Lamport et al., 1982], Computer scientists contribute significantly in designing consensus protocols that can tolerate a minority of malicious agents to achieve the goals of validity and termination. [Cachin and Vukolić, 2017, Xiao et al., 2020]. However, studies on consensus protocol in computer science often ignore the incentives that could affect agents behavior. An emerging literature in game theory introduces rational agents and solution concepts to study equilibrium outcomes including social welfare of various consensus protocols [Saleh, 2021]. However, the existing studies with rational agents are limited in generalizability and is far from providing an guidance for future designs of consensus protocols. To advance the existing literature, we strive to answer three research questions (RQs):

(1) **RQ1**: How to abstract a consensus protocol as a general game environment in extensive form [Kroer and Sandholm, 2014] that consists of the game tree, the set of agents, and the payoffs?
(2) **RQ2**: how does the behavior of bounded rational agents evolves to stable equilibria that differ in desirable outcomes of validity, termination, and social welfare?
(3) **RQ3**: how can the game theoretical study of bounded rational agents guide future designs of blockchain consensus protocols?

We succesfully abstract a general Byzantine consensus protocol as a game environment, apply bounded rationality to model agent behaviors, and solve the initial conditions for three different stable equilibria. Finally, our research guide future designs of consensus protocols to achieve desirable outcomes by evaluating incentives choices.

2 RELATED LITERATURE

Our research mainly contribute to five streams of literature across disciplines: (1) Byzantine consensus protocol in *computer science*, (2) Game Theory in *economics* on blockchain consensus,(3) Evolutionary Game theory at the intersection of *biology* and *economics*, and (4) bounded-rationality at the interplay between *psychology* and *economics*; (5) cooperative AI at the interface of *computer science* and *economics*.

2.1 Byzantine Consensus protocols

Proposed by Lamport et al. [1982], the Byzantine generals problem describes the network failure in a distributed system caused by malicious agents or corrupt nodes. Many Byzantine Fault Tolerance (BFT) protocols have been developed so far (i.e. Zyzzyva [Kotla et al., 2007], RBFT [Aublin et al., 2013], and MinBFT [Veronese et al., 2013]), among which the Practical Byzantine Fault Tolerance (PBFT) has become one of the most popular protocols. By running Byzantine consensus protocols, the distributed systems is able to tolerate a proportion of existed Byzantine agents and still achieve consensus. Compared with other non-BFT protocols (i.e. Proof-of-Work (PoW) [Nakamoto, 2008], Proof-of-Stake (PoS) [Saleh, 2018]), PBFT and its variants ([Tong et al., 2019, Wang et al., 2021]) with

30% fault tolerance has been proved to have superiority confronting the Byzantine general problems. Computer scientists endeavor to improve a consensus protocol by engineering approaches that, for example, reduce the cost and simplify the replicas in the state machine [Tong et al., 2019] or optimize the efficiency in communications [Wang et al., 2021]. In contrast, our research provide an evolutionary game approach to evaluate the consensus process for agents driven by incentives. Moreover, our results serve as a guidance to optimize the incentive scheme in consensus protocols to achieve desirable outcome.

2.2 Game Theory on Blockchain Consensus

Most of the consensus protocols are built under the assumption that non-Byzantine agents would act honestly. However, agents might be driven by incentives, which is important for protocol performance but ignored when designing the protocol. Table 1 represents recent game theory studies on blockchain consensus considering the effect of incentives on miner behaviors in the facets of consensus protocols, game theory solution concepts, agent types, and aspects for evaluations. In a hybrid consensus design, Pass and Shi [2017] propose "committees" to execute permissioned consensus protocols on permissionless blockchains.[1] Abraham et al. [2016] firstly mention incentives in the design of consensus protocols but did not provide a formal game theory study to evaluate the performance. Amoussou-Guenou et al. [2020a], Biais et al. [2019], Halaburda et al. [2021], Saleh [2018] abstract the consensus protocol to dynamic game of imperfect information [Fudenberg and Tirole, 1991a] and further introduce rational agent, game theory solution concept, and welfare analysis to study a variety of consensus protocols including the PoS [Saleh, 2018], the PoW [Biais et al., 2019], and the PBFT variants [Amoussou-Guenou et al., 2020a, Biais et al., 2019, Halaburda et al., 2021]. Above studies focus on the solution concepts of Perfect Bayesian Equilibrium (PBE, [Fudenberg and Tirole, 1991b]) or its refinement, Markov Perfect Equilibrium (MPE, [Maskin and Tirole, 2001b]. Our research contributes to existing literature by introducing bounded rational agents and the evaluation of Equilibrium Stable Strategy (ESS, [Smith and Price, 1973, Smith, 1979].

2.3 Evolutionary Game Theory

Emerged from Darwin [1859]'s evolutionary study, evolutionary models have been applied widely in many interdisciplinary studies. In *macroeconomics*, the evolutionary models were derived from the debate of Homo Economics and irrational human to explain complex phenomenon in financial markets [Henrich et al., 2001, Levin and Lo, 2021]. And in *computer science*, the evolutionary models are applied to solve graphic problems and estimate computational complexity [Nisan, 2007, Shoham and Leyton-Brown, 2009]. In game theory, Prisoner's Dilemma as the most classic example, has been studied with multiple evolutionary game designs [Axelrod and Hamilton, 1981, Shoham and Leyton-Brown, 2009, Yang and Yue, 2019], whose steady state has been proved to be an equilibrium convergence [Nisan, 2007, Shoham and Leyton-Brown, 2009] and is resilient to small mutual invasions [Shoham and Leyton-Brown, 2009]. Researchers have been applying the evolutionary game theory to study cooperation mechanism for a long time [Axelrod and Hamilton, 1981, Shoham and Leyton-Brown, 2009]. Our research advances the literature by extending the application scenario of evolutionary game theory to blockchain consensus protocols, especially in cooperation among agents.

[1] Committee-based blockchains have superiorities in reducing fork risk [Amoussou-Guenou et al., 2020a, Biais et al., 2019], expanding larger scalability [TRON, 2018], and performing robustness and efficiency [Auer et al., 2021, Benhaim et al., 2021].

Table 1. Game theory studies on blockchain consensus: dynamic game of imperfect information [Fudenberg and Tirole, 1991a]

Literature	Consensus protocol	Game Theory Solution Concept	Agent Types	Evaluation
Abraham et al. [2016]	PoW (election) PBFT variant (transaction)	none	honest or byzantine	agreement safty and liveness
Saleh [2018]	PoS (forkable)	PBE Fudenberg and Tirole [1991b]	rational	consensus social welfare
Biais et al. [2019]	PoW (forkable)	MPE Maskin and Tirole [2001b]	rational	consensus social welfare
Amoussou-Guenou et al. [2020b]	PBFT variant (non-forkable)	PBE	rational or byzantine	consensus termination validity
Halaburda et al. [2021]	PBFT variant (non-forkable)	PBE	rational or byzantine	consensus social welfare
Ours	pBFT variant (non-forkable)	ESS Smith [1979] Smith and Price [1973]	bounded rational	safety liveness social welfare

2.4 Bounded Rationality

Another core assumption of our approach is the bounded rationality of agents [Chase and Simon, 1973, Rubinstein, 2002, Simon, 1955] such that rationality is limited when making decisions. Agents in our model are bounded-rational in three facets:

(1) agents are constrained to choose among a limited set of strategies, i.e., either honest strategy or byzantine strategy;
(2) in each round of the game, agents choose strategies based on only the current state without any foresight of the future state [Conlisk, 1996].
(3) agents are allowed to hold the (inconsistent) subjective belief that they would meet agents of the same strategy with a higher probability. This assumption is supported by research in bounded rationality and game theory [Levin and Zhang, 2020, Rubinstein and Salant, 2016], the false consensus effect in psychology [Ross et al., 1977], and the assortative matching literature in behavioral science [Angelucci and Bennett, 2017, Durlauf and Seshadri, 2003, Eeckhout and Kircher, 2018, Eshel and Cavalli-Sforza, 1982, Shimer and Smith, 2000, Yang and Yue, 2019].

Our research extends bounded rationality to study the dynamics of miner behaviors in blockchain consensus. Classic game theory in *economics* assumes full rationality. In contract, consensus protocol studies in *computer science* ignores rationality. Following Aristotle's doctrine of the mean, ours is a middle way to bridge the two.

2.5 Cooperative AI

Cooperative AI studies how Artificial Intelligence (AI) can contribute to solve cooperative problems in AI-AI, AI-Human, and Human-Human interactions [Dafoe et al., 2021, 2020]. Blockchain is also coined decentralized AI [Marwala and Xing, 2018, Salah et al., 2019, Sen, 2013, Wang and Singh,

2007, Xing and Marwala, 2018]. Our research thus extend the application scenario of cooperative AI to the design of blockchain consensus.

3 MODEL

Our model is based on a BFT consensus-based blockchain that we adapted from Manshaei et al. [2018] where agents form different parallel committees to validate non-intersecting transactions. In our model, we only focus inside of one miner committee and play an evolutionary game where a mining committee participates in an n round mining game and the bounded-rational miners choose their strategies before each round of the game. To clarify, all the miners in our model has the same presentation of the agents in our paper.

3.1 Game Environment

We follow a general byzantine consensus model mentioned in Amoussou-Guenou et al. [2020b], Manshaei et al. [2018], where a PBFT consensus protocol is run. To facilitate interdisciplinary conversations, we further abstract the consensus protocol as a game environment. The environment consists of the game tree, the set of agents, and the payoffs as in Fig. 1.

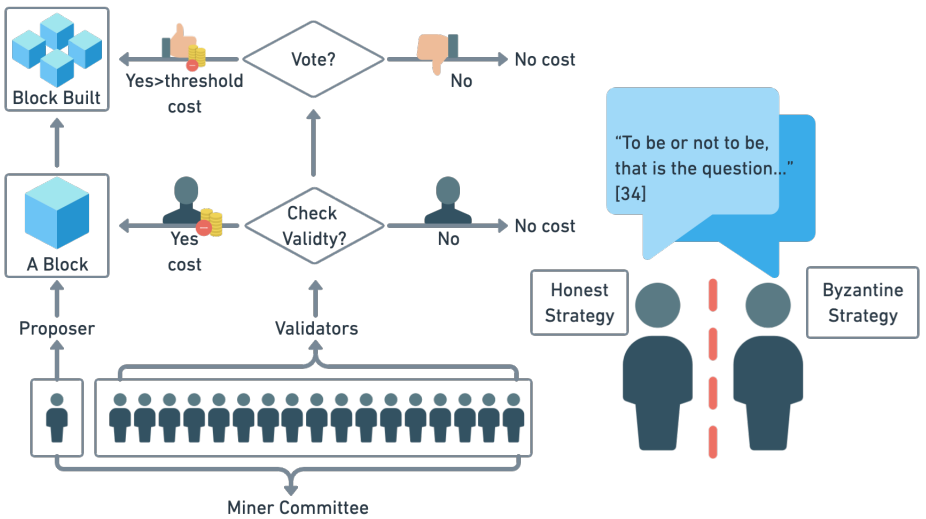

Fig. 1. Byzantine consensus-based blockchain model.

3.1.1 *Set of agents.*

Definition 1 (Miner Committee [Amoussou-Guenou et al., 2020b]). *We use an ordered set $\mathcal{A} = \{A_i\}_{i=1}^{N}$ to denote the committee of N miners established at time $t = 0$. The committee set will not change until the end of the game.*

Definition 2 (Proposer). *In round t, $t \in \{1, \cdots, n\}$, an agent A_i is selected in a round-robin fashion to be the proposer P_t. P_t makes a proposal h_t which contains a block and its validity.*

Definition 3 (Validator [Amoussou-Guenou et al., 2020b]). *In round t, $t \in \{1, \cdots, n\}$, agents except for P_t are validators. The validators are to check the validity of the proposal and vote for it when the block in the proposal is valid.*

3.1.2 Game tree.
In a miner committee $\mathcal{A}$, in round t, a proposer P_t is firstly select in a round-robin fashion and she proposes a proposal h_t which consists of a block and its validity value. The rest miners in the committee, acting as validators, then choose to check the validity of h_t and vote for it. The proposal would be accepted if the number of votes exceeds the majority threshold ν in a fixed time period. The game tree of is shown in Fig. 2 (since the final payoffs are relevant to the proportion of two strategies, Fig. 2 only shows the cost of actions). Specifically, we use

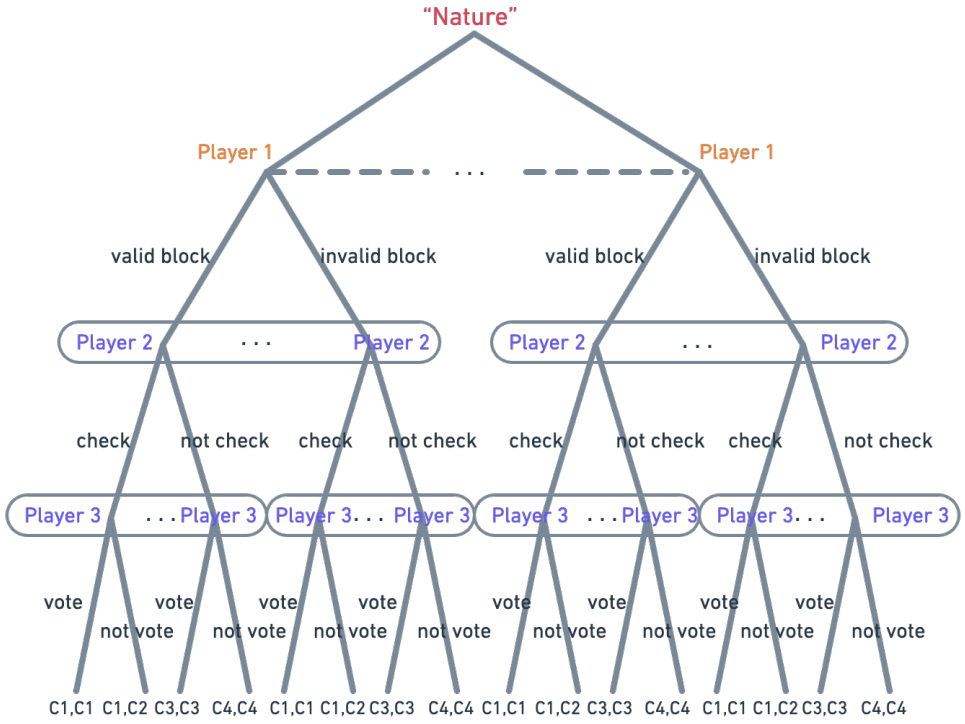

Fig. 2. Game Tree. (notations: player 1: proposer; player 2, player 3: validator; $C1 = -$(cost of check + cost of vote); $C2 = -$cost of check; $C3 = -$cost of vote; $C4 = 0$.)

3.1.3 Payoffs.
Based on [Amoussou-Guenou et al., 2020b], the payoffs of the agents include R, the reward to the validators who send a message when the block is accepted; c_{send}, the cost to the validators who send a message; c_{check}, the cost to the validators who check the validity of the proposed block; and κ, the cost occurs to all validators with the Honest strategy when an invalid block is accepted.[2]

3.2 Bounded-rational Agents

We model the miners as bounded rational agents in an evolutionary game who:

(1) in each round, only choose between honest strategy and byzantine strategy;

[2]In Amoussou-Guenou et al. [2020b], they assume $\kappa > R > c_{\text{check}} > c_{\text{send}}$.

(2) hold subjective beliefs;
(3) follow an imitative updating rule.

3.2.1 Strategies.
In round t, each of the N miners choose a strategy $s_i \in \{S_H, S_B\}$, where S_H is the honest strategy and S_B is the byzantine strategy. To explicitly define the two strategies with the bounded-rational agents, we suppose that before round t starts, each validator checks if the group of their congeners has pivotality, which means if the number of their congeners exceeds the majority threshold v.

Definition 4 (Honest Strategy). *We use S_H to denote the honest strategy, which asks a miner to achieve the consensus protocol. When playing the honest strategy:*

(1) *a proposer proposes a valid proposal (Fig. 3);*
(2) *if with pivotality, a validator checks the proposal's validity and vote for it if the proposal is valid (Fig. 4);*
(3) *if without pivotality, a validator neither checks the proposal's validity nor votes for it (Fig. 4, Fig. 5).*

Definition 5 (Byzantine Strategy). *We use S_B to denote the Byzantine strategy, which asks a miner to damage the consensus protocol. When playing the Byzantine strategy:*

(1) *a proposer proposes an invalid proposal (Fig. 3);*
(2) *if with pivotality, a validator checks the proposal's validity and vote for it if the proposal is invalid (Fig. 4);*
(3) *if without pivotality, a validator neither checks the proposal's validity nor votes for it (Fig. 4, Fig. 5).*

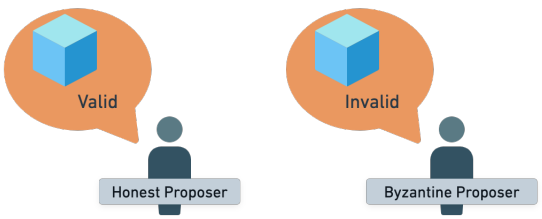

Fig. 3. Proposer actions with different strategies.

Assumption 1 (Game Initials). *We use x_t to denote the proportion of miners with S_H in round t. We assume that x_1 of the miners choose the honest strategy in the initial round.*

3.2.2 Subjective Belief.
Inspired by the Assortative Matching theory [Angelucci and Bennett, 2017, Durlauf and Seshadri, 2003, Eeckhout and Kircher, 2018, Eshel and Cavalli-Sforza, 1982, Shimer and Smith, 2000, Yang and Yue, 2019] from the Evolutionary Stable Strategy (ESS) [SMITH and PRICE, 1973], we make our assumption 2 about the validators' subjective belief to meet a proposer with the same strategy.

Assumption 2 (Subjective Belief). *We use m, $m \in [0, 1]$ to represent the portion of the rounds that a validator believes to meet a proposer with the same strategy in the game. When $m = 1$, a validator believes that in arbitrary round t, $t \in \{1, \cdots, n\}$, the proposer P_t plays the same strategy as*

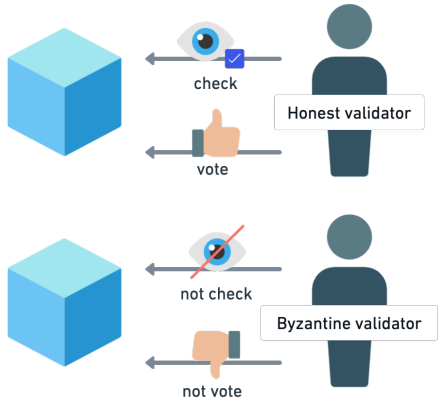

Fig. 4. Validator actions with pivotality.

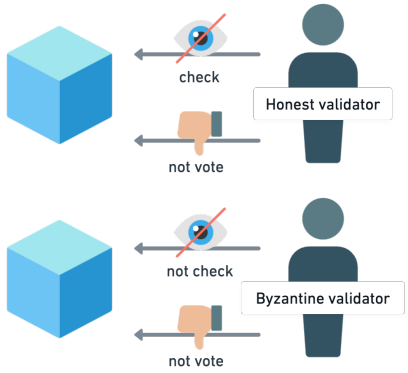

Fig. 5. Validator actions without pivotality.

she does; When $m = 0$, a validator believes that in arbitrary round t, $t \in \{1, \cdots, n\}$, the proposer P_t plays the different strategy from what she does.

Assumption 2 is also consistent with the False Consensus Effect [Ross et al., 1977] as $m \geqslant 0$. Therefore, the validators' expected payoff will be updated according to our assumption 2.

Lemma 1 (Subjective Meeting Probabilities). *We use $\pi_{ij}(x_t)$, $i, j \in \{H, B\}$ to denote the subjective meeting probability of one validator with strategy S_i meet one proposer with strategy S_j in function of x_t.*

$$\begin{aligned} \pi_{HH}(x_t) &= m + (1-m)x_t \\ \pi_{HB}(x_t) &= (1-m)(1-x_t) \\ \pi_{BH}(x_t) &= (1-m)x_t \\ \pi_{BB}(x_t) &= 1-(1-m)x_t \end{aligned} \tag{1}$$

Lemma 2 (Agents Expected Payoff). *We use $V_{ij}(x_t)$ to denote the expected payoff that one validator with S_i meets a proposer with S_B in function of x_t, where $i, j \in \{H, B\}$. We also use $V_i(x_t)$, $i \in \{H, B\}$ to denote the expected validator payoff with subjective belief of one validator with strategy S_i in function of x_t.*

$$\begin{aligned} V_H(x_t) &= \pi_{HB}(x_t)V_{HB}(x_t) + \pi_{HH}(x_t)V_{HH}(x_i) \\ V_B(x_t) &= \pi_{BB}(x_t)V_{BB}(x_t) + \pi_{BH}(x_t)V_{BH}(x_i) \end{aligned} \tag{2}$$

Moreover, as validators behaviors are crucial to the consensus building and evolutionary game updates, we treat the proposers' expected payoff the same as the validators' payoff as long as they play the same strategy.

3.2.3 *Imitative Updating Rule.*

Assumption 3 (Imitative Updating Rule[Amoussou-Guenou et al., 2020b]). *Before round t, for one individual bounded-rational miner, she has P_{H_t} probability to choose the honest strategy in round t and P_{B_t} probability to choose the Byzantine strategy in round t.*

$$P_{H_t} = \frac{x_{t-1}V_H(x_{t-1})}{x_{t-1}V_H(x_{t-1}) + (1-x_{t-1})V_B(x_{t-1})}$$

$$P_{B_t} = \frac{(1 - x_{t-1})V_B(x_{t-1})}{x_{t-1}V_H(x_{t-1}) + (1 - x_{t-1})V_B(x_{t-1})}$$

Therefore, by summing up the individual validators' strategic choices, we can calculate the proportion of validators playing the honest strategy in round t is:

$$x_t = \frac{x_{t-1}V_H(x_{t-1})}{x_{t-1}V_H(x_{t-1}) + (1 - x_{t-1})V_B(x_{t-1})}$$

3.3 Stable equilibrium

3.3.1 Definitions.
We use the Evolutionary Stable Strategy (ESS) to define the stable equilibria in our model and also apply Steady-State Equilibrium to explain the results.

As proposed by SMITH and PRICE [1973] and formalized by Eshel and Cavalli-Sforza [1982] and McKenzie [2009], the ESS describes the dominance and stability of one strategy in the system even with a few invaders with other strategies. In our model, similar to the form in Eshel and Cavalli-Sforza [1982], the stable equilibrium is reached when:

$$x_t = x_{t-1}, \textit{where } x_t = \frac{x_{t-1}V_H(x_{t-1})}{x_{t-1}V_H(x_{t-1}) + (1 - x_{t-1})V_B(x_{t-1})}. \tag{3}$$

Equation (3) results in three equilibria with different dominant strategy:

(1) Stable equilibrium 1: $x_t = x_{t-1} = 1$, S_H becomes the dominant strategy;
(2) Stable equilibrium 2: $x_t = x_{t-1} = 0$, S_B becomes the dominant strategy;
(3) Stable equilibrium 3: $V_H = V_B$, $x_t = x_{t-1} \in (0, 1)$, the system reaches a steady-state equilibrium [citation], representing that both strategies exist with a stable proportion because the numbers of agents changing between the two strategies equal.

To evaluate the validity of the steady state, we further define three types of stable equilibria as stable Byzantine-free, Byzantine-pooling, and Byzantine equilibrium.

Definition 3.1 (Stable Byzantine-free Equilibrium). We define that the stable Byzantine-free equilibrium in our model is reached when the ESS is reached and none of the agents is playing the Byzantine strategy, which mathematically represents:

$$x_{t-1} = x_t = 1$$

This also satisfies the ESS in the form of McKenzie [2009]'s work by:

$$\pi(S_H|S_H) > \pi(S_B|S_H)^3$$

Definition 3.2 (Stable Byzantine Equilibrium). We define that the stable Byzantine equilibrium in our model is reached when the ESS is reached and all of the agents are playing the Byzantine strategy, which mathematically represents:

$$x_{t-1} = x_t = 0$$

This also satisfies the ESS in the form of McKenzie [2009]'s work by:

$$\pi(S_B|S_B) > \pi(S_H|S_B)$$

[3]In McKenzie [2009], $\pi(x|y)$ denotes the payoff an agent obtained when playing strategy x against someone using the strategy y.

Definition 3.3 (Stable Byzantine-pooling Equilibrium). We define that the stable Byzantine-pooling Equilibrium in our model is reached when the stable equilibrium is reached and both of the two strategies exist, which mathematically represents:

$$x_{t-1} = x_t \in (0, 1)$$

This also satisfies the Steady-State equilibrium in the sense that the number of agents changing from S_H to S_B is equal to the number of agents changing from S_B to S_H. As defined by Gagniuc [2017], a steady-state equilibrium is reached when the variables defining the behavior of the system or the process are unchanging in time. In our stable Byzantine-pooling equilibrium, the numbers of the agents playing the two strategies are unchanging in the game.

3.3.2 *Propositions.*

With different initial values of x_1 and R, c_{check}, c_{send}, κ, and ν, we can now have the following propositions as our results.

Proposition 1. *The stable Byzantine-free equilibrium can be reached if and only if:*

$$1 - \frac{\nu}{N} \geqslant x_1 > \max\left(\frac{R - c_{check} + \kappa}{2R - 2c_{check} + \kappa}, \frac{\nu}{N}\right) \quad and \quad m \neq 1, \quad or \quad x_1 > \max\left(1 - \frac{\nu}{N}, \frac{\nu}{N}\right)$$

is satisfied.

Lemma 3. *If initially $N(1 - x_1) \geqslant \nu$, meaning the initial number of agents with Byzantine strategy is equal or larger than the majority threshold, only if $x_1 > \max(\frac{R - c_{check} + \kappa}{2R - 2c_{check} + \kappa}, \frac{\nu}{N})$ and $m \neq 1$ the stable Byzantine-free equilibrium can be reached.*

Lemma 4. *If initially $N(1 - x_1) < \nu$, which represents that the initial number of agents with Byzantine strategy is smaller than the threshold, only if when $x_1 \geqslant \frac{\nu}{N}$, meaning that the initial number of agents with the honest strategy is larger than the threshold, the stable Byzantine-free equilibrium can be reached*

Proof for Lemma 3. When $N(1-x_1) \geqslant \nu$, agents playing S_B has pivotality in the voting process. Therefore, to reach the stable Byzantine-free equilirbium, the initial number of agents playing S_H should at least reach ν, meaning $Nx_1 \geqslant \nu$.

Validator behaviors and proposal outcome In round 1, if the proposer P_1 plays S_B, she proposes an invalid proposal h_1. After all the agents check the validity of h_t, agents with S_B vote for h_1 while agents with S_H do nothing. h_1 will be accepted since $N(1 - x_1) \geqslant \nu$.

Therefore, the payoffs in round 1 for the validators with P_1 playing S_B are:

$$\begin{aligned} V_{BB}(x_1) &= R - c_{check} - c_{send} \\ V_{HB}(x_1) &= -c_{check} - \kappa \end{aligned} \tag{4}$$

In round 1, if the proposer P_1 plays S_H, she proposes a valid proposal h_1. After all the agents check the validity of h_1, agents with S_H vote for h_1 while agents with S_B do nothing. h_1 will be accepted since $Nx_1 \geqslant \nu$.

Therefore, the payoffs in round 1 for the validators with P_1 playing S_H are:

$$\begin{aligned} V_{HH}(x_1) &= R - c_{check} - c_{send} \\ V_{BH}(x_1) &= -c_{check} \end{aligned} \tag{5}$$

Validator Payoff matrix The payoff matrix for validators in round 1 when initially $n(1 - x_1) \geqslant \nu$ and $nx_1 \geqslant \nu$ is shown in Table 2.

Table 2. Payoff matrix when $n(1-x_1) \geqslant \nu$ and $nx_1 \geqslant \nu$

	Proposer – Honest	Proposer – Byzantine
Validator – Honest	$V_{HH}(x_1) = R - c_{check} - c_{send}$	$V_{HB}(x_1) = -c_{check} - \kappa$
Validator – Byzantine	$V_{BH}(x_1) = -c_{check}$	$V_{BB}(x_1) = R - c_{check} - c_{send}$

From equations (1) from lemma 1, (2) from lemma 2, (4), and (5), we can get:

$$\begin{aligned} V_H(x_1) &= [m + (1-m)x_1](R - c_{check} - c_{send}) + (1-m)(1-x_1)(-c_{check} - \kappa) \\ V_B(x_1) &= x_1(1-m)(-c_{check}) + [m + (1-m)(1-x_1)](R - c_{check} - c_{send}) \end{aligned} \tag{6}$$

To achieve the stable Byzantine-free equilibrium, x_t should converge to 1. Then for round t and $t-1$ ($t \geqslant 2$) before convergence,

$$x_t = \frac{x_t V_H(x_{t-1})}{x_t V_H(x_{t-1}) + (1-x_t)V_B(x_{t-1})} > x_{t-1} \tag{7}$$

Formula (7) can be simplified to (since $x_t \neq x_{t-1}$ and $x_{t-1} \in (0,1)$):

$$V_H(x_{t-1}) > V_B(x_{t-1}) \tag{8}$$

And by combining equation (6) with (8) and deduction with $t = 2$:

$$V_H(x_1) - V_B(x_1) = -(1-m)[(1-2x_1)(R - c_{send}) + (1-x_1)\kappa] \tag{9}$$

Equation (9) implies that, if $m \neq 1$ and $x > \frac{R - c_{check} + \kappa}{2R - 2c_{check} + \kappa}$, x_t would converge to 1, and the system will reach the stable Byzantine-free equilibrium. The backward proof can be implemented by taking $x_1 > \max(\frac{R - c_{check} + \kappa}{2R - 2c_{check} + \kappa}, \frac{\nu}{N})$ and $m \neq 1$ into formula (8). Therefore, when $N(1-x_1) \geqslant \nu$, the stable Byzantine-free equilibrium can be reached if and only if $x_1 > \max(\frac{R - c_{check} + \kappa}{2R - 2c_{check} + \kappa}, \frac{\nu}{N})$ and $m \neq 1$.

□

Proof for Lemma 4. When $N(1-x_1) < \nu$, agents playing S_B do not have pivotality in the voting process. Therefore, to reach the stable Byzantine-free equilibrium, the initial number of agents playing S_H should at least reach ν, meaning $Nx_1 \geqslant \nu$.

Validator behaviors and proposal outcome In round 1, if the proposer P_1 plays S_B, she proposes an invalid proposal h_1. Agents with S_H would check the validity of h_t but not vote for it, while agents with S_B do nothing because they do not have pivotality in the voting process. h_1 will not be accepted since no one votes for it.

Therefore, the payoffs in round 1 for the validators with P_1 playing S_B are:

$$\begin{aligned} V_{BB}(x_1) &= 0 \\ V_{HB}(x_1) &= -c_{check} \end{aligned} \tag{10}$$

In round 1, if the proposer P_1 plays S_H, she proposes a valid proposal h_1. Agents with S_H would check the validity of h_t and vote for it, while agents with S_B do nothing because they do not have pivotality in the voting process. h_1 will be accepted since $Nx_1 \geqslant \nu$.

Therefore, the payoffs in round 1 for the validators with P_1 playing S_H are:

$$\begin{aligned} V_{HH}(x_1) &= R - c_{check} - c_{send} \\ V_{BH}(x_1) &= 0 \end{aligned} \tag{11}$$

Validator Payoff matrix The payoff matrix for validators in round 1 when initially $n(1-x_1) < \nu$ and $nx_1 \geqslant \nu$ is shown in Table 3.

Table 3. Payoff matrix when $n(1-x_1) < v$ and $nx_1 > v$

	Proposer – Honest	Proposer – Byzantine
Validator – Honest	$V_{HH}(x_1) = R - c_{check} - c_{send}$	$V_{HB}(x_1) = -c_{check}$
Validator – Byzantine	$V_{BH}(x_1) = 0$	$V_{BB}(x_1) = 0$

From equations (1) from lemma 1, (2) from lemma 2, (10), and (11), we can get:

$$\begin{aligned} V_H(x_1) &= [m + (1-m)x_1](R - c_{check} - c_{send}) + (1-m)(1-x_1)(-c_{check}) \\ V_B(x_1) &= 0 \end{aligned} \tag{12}$$

Similar as we have proved in proof of lemma 3, formula (8) is satisfied with equation (12) with $t = 2$. And the backward proof can be implemented by taking $x_1 \geqslant \frac{v}{N}$ with formula (8) and equation (12). Therefore, when $N(1-x_1) < v$, the stable Byzantine-free equilibrium equilibrium can be reached if and only if $x_1 \geqslant \frac{v}{N}$.

□

Proposition 2. *The Byzantine equilibrium can be reached if and only if:*

$$\frac{v}{N} \leqslant x_1 < min(\frac{R - c_{check} + \kappa}{2R - 2c_{check} + \kappa}, 1 - \frac{v}{n}) \quad and \quad m \neq 1, \quad or \quad x_1 < \min(1 - \frac{v}{N}, \frac{v}{N}) \tag{13}$$

is satisfied.

Lemma 5. *If initially $N(1-x_1) \geqslant v$, meaning the initial number of agents with Byzantine strategy is equal or larger than the majority threshold, only if $x_1 < \min(\frac{R-c_{check}+\kappa}{2R-2c_{check}+\kappa}, \frac{v}{N})$ and $m \neq 1$, or $x < min(1 - \frac{v}{N}, \frac{v}{N})$ the stable Byzantine equilibrium can be reached.*

Proof for Lemma 5. The proof for stable Byzantine equilibrium achievement when $x_1 < \max(\frac{R-c_{check}+\kappa}{2R-2c_{check}+\kappa}, \frac{v}{N})$ and $m \neq 1$ is similar to the proofs of lemma 3 by changing formula (7) into:

$$x_t = \frac{x_t V_H(x_{t-1})}{x_t V_H(x_{t-1}) + (1-x_t)V_B(x_{t-1})} < x_{t-1} \tag{14}$$

And the proof for stable Byzantine equilibrium achievement when $x < min(1 - \frac{v}{N}, \frac{v}{N})$ is as below:

In contract with the assumption that $Nx_1 \geqslant v$, when $Nx_1 < v$, the agents with S_H do not have pivotality in the voting process.

Validator behaviors and proposal outcome In round 1, if the proposer P_1 plays S_B, she proposes an invalid proposal h_1. Agents with S_B would check the validity of h_1 and vote for it, while agents with S_H do nothing because they don't have pivotality in the voting process. h_1 will be accepted since $N(1-x_1) \geqslant v$.

Therefore, the payoffs in round 1 for the validators with P_1 playing S_B are:

$$\begin{aligned} V_{BB}(x_1) &= R - c_{check} - c_{send} \\ V_{HB}(x_1) &= -\kappa \end{aligned} \tag{15}$$

In round 1, if the proposer P_1 plays S_H, she proposes a valid proposal h_1. Agents with S_B would check the validity of h_1 but not vote for it, while agents with S_H do nothing because they don't have pivotality in the voting process. h_1 will not be accepted since no one vote for it.

Therefore, the payoffs in round 1 for the validators with P_1 playing S_H are:

$$\begin{aligned} V_{HH}(x_1) &= 0 \\ V_{BH}(x_1) &= -c_{check} \end{aligned} \tag{16}$$

Validator Payoff matrix The payoff matrix for validators in round 1 when initially $n(1-x_1) \geqslant v$ and $nx_1 < v$ is shown in Table 5.

Table 4. Payoff matrix when $n(1-x_1) \geqslant v$ and $nx_1 < v$

	Proposer – Honest	Proposer – Byzantine
Validator – Honest	$V_{HH}(x_1) = 0$	$V_{HB}(x_1) = -\kappa$
Validator – Byzantine	$V_{BH}(x_1) = -c_{check}$	$V_{BB}(x_1) = R - c_{check} - c_{send}$

From equations (1) from lemma 1, (2) from lemma 2, (4), and (5), we can get:

$$\begin{aligned} V_H(x_1) &= (1-m)(1-x_1)(-\kappa) \\ V_B(x_1) &= x_1(1-m)(-c_{check}) + [m + (1-m)(1-x_1)](R - c_{check} - c_{send}) \end{aligned} \tag{17}$$

Equation (17) shows that, the expected payoff of a validator with S_H is smaller than a validator with S_B, and since $n(1-x_1) \geqslant v$ and $nx_1 < v$, $1 - x_1 > x_1$. Therefore,

$$x_t = \frac{x_t V_H(x_{t-1})}{x_t V_H(x_{t-1}) + (1-x_t)V_B(x_{t-1})} > x_{t-1} \tag{18}$$

is achieved with any t before convergence. The system will converge to Byzantine equilibrium. The backward proof can be implemented by taking the corresponding values of x_1 into equations (14) and (17). Therefore, if $N(1-x_1) \geqslant v$, only if $x_1 < \min(\frac{R-c_{check}+\kappa}{2R-2c_{check}+\kappa}, \frac{v}{N})$ and $m \neq 1$, or $x < min(1-\frac{v}{N}, \frac{v}{N})$ the stable Byzantine equilibrium can be reached.

□

Proposition 3. *The Byzantine pooling equilibrium can be reached if and only if:*

$$1 - \frac{v}{N} \geqslant x_1 \geqslant \frac{v}{N}, \quad \textit{and} \quad x_1 = \frac{R - c_{check} + \kappa}{2R - 2c_{check} + \kappa} \quad \textit{or} \quad m = 1 \tag{19}$$

is satisfied.

Proof for Proposition 3. In the proof for lemma 3, equation (9) implies that, if $m = 1$ or $x = fracR - c_{check} + \kappa 2R - 2c_{check} + \kappa$, $V_H(x_1) = V_B(x_1)$, and therefore equation (7) would become $x_t = \frac{x_t V_H(x_{t-1})}{x_t V_H(x_{t-1}) + (1-x_t)V_B(x_{t-1})} = x_{t-1}$. Since $x_t = x_{t-1} \in (0,1)$, the stable Byzantine-pooling equilibrium is reached. Proof for lemma shows that the stable Byzantine-pooling equilibirum is never reached when initially $N(1-x_1) < v$ and $x_1 \geqslant v$. The backward proof can be implemented by taking $m = 1$ or $x = fracR - c_{check} + \kappa 2R - 2c_{check} + \kappa$ into formula (7) and it would become an equation.

□

3.3.3 *Evaluations.*

We evaluate our model from four different perspectives: blockchain consensus enhancement, false consensus effect on behavioral agents, the economic representations of parameters in our model, and the applicability of our model in real-life applications.

From the perspective of blockchain consensus enhancement, we would use **safety** and *liveness* as our evaluation criteria.

Safety Defined by Castro and Liskov [2002], the safety property of a BFT protocol should have the representation in our model that all the agents playing the honest strategy vote for the same block, which has been achieved by our agent strategy.

liveness The liveness property correspondingly in our model represents the block building is not stopped and the bounded-rational agents are changing their strategies according to the imitative

updating rule. There is only one scenario having liveness failure, which is not included in our stable equilibria (Lemma 5).

Lemma 6. *If initially $n(1-x_1) < v$, which represents that the initial number of agents with Byzantine strategy is smaller than the threshold, only if when $nx_1 < v$, meaning that the initial number of agents with the honest strategy is also smaller than the threshold, the protocol will never converge to one single strategy and the protocol liveness cannot be guaranteed.*

Proof. Since neither the number of miners with the honest strategy or the number of miners with the Byzantine strategy is larger than the threshold, no block will be built, and none of the bounded-rational miners will check the validity nor send a message. Therefore, the payoff for all two strategies is zero and the block-building process is stopped thus the protocol meets a failure. □

Moreover, with the evolutionary model, our model enhance the original consensus achieved by the PBFT protocol Castro and Liskov [2002] by incentivizing the agents to transferring to the honest strategy. While the most most common situation is when the initial numbers of the two strategies both exceed the majority threshold, which is the same as shown in lemma 3, proposition 1 shows that, blockchain consensus enhancement can be achieved. And our result shows that, when evaluated by protocol safety and liveness, our model can achieve consensus with the initial proportion of agents playing the honest strategy

$$1-\frac{v}{N} \geqslant x_1 \geqslant \max\left(\frac{1}{2}+\frac{\frac{1}{2}\kappa}{2R-2c_{check}+\kappa}, \frac{v}{N}\right) \quad \text{or} \quad x_1 > \max\left(1-\frac{v}{N}, \frac{v}{N}\right)$$

, which is looser than the requirements by a single PBFT protocol.

From the perspective of behavior science, our model shows how the false consensus effect influences the behavioral agents and the outcome. We model the agents' false consensus by adding a parameter m, the proportion of the rounds that a validator believes to meet a proposer with the same strategy in the game. And as our stable Byzantine-pooling equilibrium shows that, a very strong of false consensus of the behavioral agents has a great effect in making the expected payoffs of the two strategies the same in agents belief. With the false consensus, the agents are over-confident with their choice and act bounded-rationally.

From the economic representation perspective in our model, we firstly reparametrize the stable equilibria conditions: Let $\alpha = \frac{R}{\kappa}$, $\beta = \frac{c_{\text{check}}}{\kappa}$, and $\gamma = \frac{v}{N}$. By proposition 1—3, the range of x_1 that achieves each equilibrium only depends on γ, α, and β (Table. 5).

Table 5. Stable equilibria with reparametrization

Equilibrium Type	Conditions	Agents Payoff
Stable Byzantine-free Equilibrium	$1-\gamma \geqslant x_1 > \max(\frac{1}{2}+\frac{\frac{1}{2}}{2\alpha-2\beta+1}, \gamma)$ or $x_1 > \max(1-\gamma, \gamma)$	Equation (6) Equation (12)
Stable Byzantine Equilibrium	$\gamma \leqslant x_1 < min(\frac{1}{2}+\frac{\frac{1}{2}}{2\alpha-2\beta+1}, 1-\gamma)$ or $x_1 < \min(1-\gamma, \gamma)$	Equation (6) Equation (17)
Stable Byzantine-pooling Equilibrium	$1-\gamma \geqslant x_1 = \frac{1}{2}+\frac{\frac{1}{2}}{2\alpha-2\beta+1} \geqslant \gamma$	Equation (6)

We conduct some numerical analysis to evaluate how the change of parameters would influence the stable equilibria achievement requirement for x_1 (Fig. 6). Specifically we focus on the changes of parameters α and β as they are related with the agents incentives. We simulate our model with $1-\gamma \geqslant x \geqslant \gamma$ as a premise. The results shows that, when the value of α or β increases, the initial

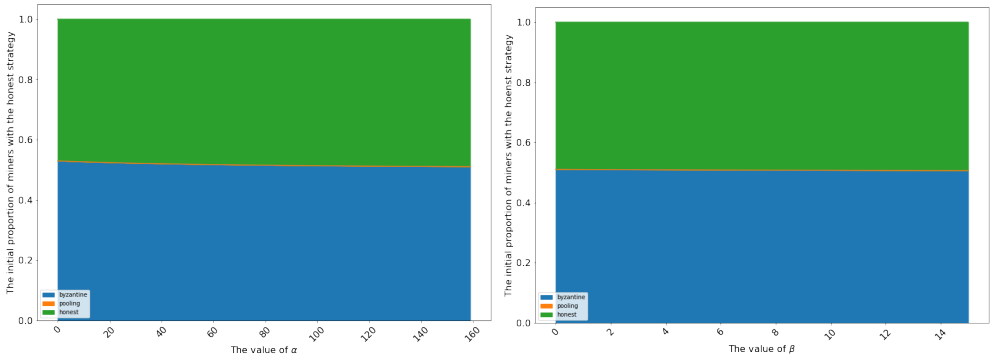

Fig. 6. Simulations for the range of x_1 that ends up with different equilibria in steady states: (parameters choice: left: $\alpha \in [4, 20]$, $\beta = 3$, with $1 - \gamma \geqslant x \geqslant \gamma$; right: $\alpha = 20$, $\beta \in [3, 19]$, with $1 - \gamma \geqslant x \geqslant \gamma$.)

proportion of miners with the honest strategy required to reach the Byzantine-free equilibrium decreases.

From the perspective of the applicability of our model in real-life applications, we show how the changes of the policy parameters in our model can be implemented in real blockchain system configuration and regulation. As Ethereum is upgrading to 2.0 version, we can take some of the key features from it as an example. Table 7 shows the parameters in our model that affect equilibrium outcomes and their corresponding policy parameters in Ethereum 2.0. Although Ethereum 2.0 will be using PoS as its consensus protocol, our model provides an approach to model the miners as bounded-rational agents and set policy parameters to incentivize them to enhance the consensus achievement on chain. Although all our three stable equilibria reach consensus, they differ in aspects of safety, liveness, and social welfare. Table 7 shows the parameters that affect equilibrium outcomes in steady states. Practitioners could experiment with the policy parameters to achieve desirable outcomes. The

Table 6. Policy parameters that affect equilibrium outcomes in steady states

notation	definition in our model	corresponding parameters in Ethereum 2.0
R	the reward to the validators who send a message when the block is accepted	Reward Per Block(about 2ETH) [BitInfoCharts, 2019]
c_{check}	the cost to the validators who check the validity of the proposed block	
c_{send}	the cost of the validators who send a message	
κ	the cost occurs to all validators with the honest strategy when an invalid block is accepted	slashing[He and Li, 2022]
ν	the majority threshold of the votes	Depositing 32 ETH to activate a validator[Foundation, 2022]
x_1	the initial number of agents with honest strategies	KYC problem [Parra Moyano and Ross, 2017]

4 DISCUSSIONS AND FUTURE RESEARCH

In this paper, we demonstrate a general Byzantine consensus protocol as a game environment, apply bounded rationality to model agent behaviors, and solves for the initial conditions for three different stable equilibria.

Fundamentally, our research suggests a new approach for consensus enhancement on a BFT blockchain with incentive-driven bounded-rational agents. Besides, our dynamic design of the evolutionary game for our model, together with the stable equilibria of ESS and Steady-state provides a new methodology to achieve blockchain consensus and helps to explore cooperative scenarios on the trust machine. Moreover, our results show high consistency with the open problems in cooperative AI and our simulation also verifies one of the hypothesis for mitigating the problems.

4.1 Blockchain consensus enhancement

This paper mainly contribute to blockchain consensus achievement by suggesting a game theory approach to enhance consensus building by incentivizing all bounded-rational agents to play the honest strategy and achieve the protocol. Among all the game theory studies on blockchain consensus, our approach is the first one that supports consensus enhancement. While being compared with other traditional BFT algorithms for consensus building, which achieves the protocol by tolerating network failure caused by faulty agents, our approach innovatively proposes the possibility in transferring the faulty agents into honest ones even when the proportion of honest strategy is slightly larger than $\frac{1}{2}$ in a common situation.

4.2 Game theoretical approach

Our research is seminal in game theoretical approach for blockchain consensus problem that can be extended in the three facets below.

(1) **game environment**: On blockchain there are a variety of consensus protocols [Cachin and Vukolić, 2017, Sankar et al., 2017, Xiao et al., 2020], each provides a game environment. Future research can follow our approach to abstract other consensus protocols to a new game environment for performance evaluation and comparative studies.
(2) **bounded rational agent**: Future research could consider including other bounded rational agents such as prominent the Level-K [Arad and Rubinstein, 2012, Crawford and Iriberri, 2007, Levin and Zhang, 2020] agent in economics and the trending reinforcement learning agent [Ivanov, 2022, Sutton and Barto, 2018] in computational science.
(3) **equilibrium solution concepts**: Future research could compare the results in a variety of game theory solution concepts [Tardos and Vazirani, 2007] and test the performance in empirical or experimental studies.

4.3 Cooperation on Blockchain

In this paper, cooperation of the miners is embodied in consensus achievement on the blockchain as well as the social welfare maximization of miners rewards. Our work shows three key features of cooperation on blockchain:

(1) **On blockchain, a trust machine, miners are originally to have commitment and work cooperative work.** This feature comes from the nature of blockchain. As a distributed system, blockchain is built by miners. While the miners are incentivized by the block rewards and future transactions on the block, their welfare is maximized when the stability of blockchain is also maximized. Any malicious behavior or system hack would cause great loss in miner rewards. Therefore, it is the nature of blockchain that miners are to have commitment and work cooperatively in block building process.

(2) **From the perspective of long-term cooperation, cooperation on blockchain consensus achievement gives the most social welfare and individual reputation for continuous work.** By constructing an evolutionary game, our paper shows the long-term cooperation of miners in block building. Similar to the analysis in Schneider and Weber [2013]'s work, our work uses the bounded-rational agents' choices of strategy to represent miner's willingness to interact with other participants and have commitment with the committee in the long-term cooperation. Consistent with prior research, our results also show that miner's willingness to longer commitment (choosing the honest strategy) will give higher social welfare when the global expected outcome is to achieve consensus and support block building.

(3) **The blockchain committees can be modelled as coalitions and competitions between can support the cooperation within each committee and therefore works for the whole blockchain building.** Since our research only focuses on consensus achievement within one committee, considering the competition effects between coalitions [Staatz, 1983], we suggest that instead of practising the same model to all parallel committees, future research should consider to analyze how competitions between miner committees support global consensus achievement.

4.4 Problems and hypothesis in cooperative AI

Our result shows the potential scenario when one of the downsides of cooperative AI reveals in cooperation on blockchain. As mentioned and explained by Dafoe et al. [2020], one the potential downsides of cooperative AI is the appearance of exclusion and collusion, which happens when the cooperation of agents harms external environment or other agents not involved by the cooperation, or when the cooperation undermines pro-social competition. In our stable equilibria analysis, we show that when the system reaches the Byzantine equilibrium, all the agents play the Byzantine strategy that harms the blockchain consensus and security. Therefore, the convergence to Byzantine equilibrium shows the potential scenario of harmful cooperation in blockchain system. To mitigate such cooperation downsides, Dafoe et al. [2020] offered the *Hypothesis that Broad Cooperative Competence is Beneficial* which we partially verifies with the model simulation. Fig. 6 shows that, when any of the value of α and β increases, which represents a larger incentive in the global environment and therefore raise the board cooperative competence, the system is able to transfer a larger proportion of faulty agents with incentives.

REFERENCES

2018. The cooperative human. 2, 7 (2018), 427–428. https://doi.org/10.1038/s41562-018-0389-1

Ittai Abraham, Dahlia Malkhi, Kartik Nayak, Ling Ren, and Alexander Spiegelman. 2016. Solidus: An Incentive-compatible Cryptocurrency Based on Permissionless Byzantine Consensus. (12 2016).

Yackolley Amoussou-Guenou, Bruno Biais, Maria Potop-Butucaru, and Sara Tucci-Piergiovanni. 2020a. Rational Behavior in Committee-Based Blockchains. https://hal.archives-ouvertes.fr/hal-02867095

Yackolley Amoussou-Guenou, Bruno Biais, Maria Potop-Butucaru, and Sara Tucci-Piergiovanni. 2020b. Rational vs Byzantine Players in Consensus-Based Blockchains. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems* (Auckland, New Zealand) *(AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 43–51.

Manuela Angelucci and Daniel Bennett. 2017. Assortative Matching under Asymmetric Information: Evidence from Malawi. *American Economic Review* 107 (05 2017), 154–157. https://doi.org/10.1257/aer.p20171055

Ayala Arad and Ariel Rubinstein. 2012. The 11-20 money request game: A level-k reasoning study. *American Economic Review* 102, 7 (2012), 3561–73.

Pierre-Louis Aublin, Sonia Ben Mokhtar, and Vivien Quema. 2013. RBFT: Redundant Byzantine Fault Tolerance. *2013 IEEE 33rd International Conference on Distributed Computing Systems* (07 2013). https://doi.org/10.1109/icdcs.2013.53

Raphael Auer, Cyril Monnet, and Hyun Song Shin. 2021. Distributed ledgers and the governance of money. *www.bis.org* (01 2021). https://www.bis.org/publ/work924.htm

R Axelrod and W. Hamilton. 1981. The evolution of cooperation. *Science* 211 (03 1981), 1390–1396. https://doi.org/10.1126/science.7466396

Alon Benhaim, Brett Hemenway Falk, and Gerry Tsoukalas. 2021. Scaling Blockchains: Can Elected Committees Help? *SSRN Electronic Journal* (2021). https://doi.org/10.2139/ssrn.3914471

Bruno Biais, Christophe Bisière, Matthieu Bouvard, and Catherine Casamatta. 2019. The Blockchain Folk Theorem. *The Review of Financial Studies* 32 (04 2019), 1662–1715. https://doi.org/10.1093/rfs/hhy095

BitInfoCharts. 2019. Ethereum / Ether (ETH) statistics - Price, Blocks Count, Difficulty, Hashrate, Value. https://bitinfocharts.com/ethereum/

Christian Cachin and Marko Vukolić. 2017. Blockchain consensus protocols in the wild. *arXiv preprint arXiv:1707.01873* (2017).

Miguel Castro and Barbara Liskov. 2002. Practical byzantine fault tolerance and proactive recovery. *ACM Transactions on Computer Systems* 20 (11 2002), 398–461. https://doi.org/10.1145/571637.571640

William G. Chase and Herbert A. Simon. 1973. Perception in chess. *Cognitive Psychology* 4 (01 1973), 55–81. https://doi.org/10.1016/0010-0285(73)90004-2

John Conlisk. 1996. Why Bounded Rationality? *Journal of Economic Literature* 34 (1996), 669–700. https://www.jstor.org/stable/2729218

Vincent P Crawford and Nagore Iriberri. 2007. Level-k auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions? *Econometrica* 75, 6 (2007), 1721–1770.

Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2021. Cooperative AI: machines must learn to find common ground.

Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. 2020. Open Problems in Cooperative AI. *arXiv:2012.08630 [cs]* (12 2020). https://arxiv.org/abs/2012.08630

Charles Darwin. 1859. *On the Origin of Species.* Natural History Museum.

Steven N. Durlauf and Ananth Seshadri. 2003. Is assortative matching efficient? *Economic Theory* 21 (03 2003), 475–493. https://doi.org/10.1007/s00199-002-0269-8

Jan Eeckhout and Philipp Kircher. 2018. Assortative Matching With Large Firms. *Econometrica* 86 (2018), 85–132. https://doi.org/10.3982/ecta14450

Ilan Eshel and L. L. Cavalli-Sforza. 1982. Assortment of Encounters and Evolution of Cooperativeness. *Proceedings of the National Academy of Sciences of the United States of America* 79 (1982), 1331–1335. https://www.jstor.org/stable/12036

Ernst Fehr and Urs Fischbacher. 2004. Social norms and human cooperation. *Trends in Cognitive Sciences* 8 (04 2004), 185–190. https://doi.org/10.1016/j.tics.2004.02.007

Ernst Fehr, Urs Fischbacher, and Simon Gächter. 2002. Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature* 13 (03 2002), 1–25. https://doi.org/10.1007/s12110-002-1012-7

Ernst Fehr and Simon Gächter. 2000. Cooperation and Punishment in Public Goods Experiments. *American Economic Review* 90 (09 2000), 980–994. https://doi.org/10.1257/aer.90.4.980

Ethereum Foundation. 2022. Solo stake your ETH. https://ethereum.org/en/staking/solo/

Drew Fudenberg and Jean Tirole. 1991a. *Game theory.* MIT press.

Drew Fudenberg and Jean Tirole. 1991b. Perfect Bayesian equilibrium and sequential equilibrium. *journal of Economic Theory* 53, 2 (1991), 236–260.

Drew Fudenberg and Jean Tirole. 1991c. Perfect Bayesian equilibrium and sequential equilibrium. *Journal of Economic Theory* 53 (04 1991), 236–260. https://doi.org/10.1016/0022-0531(91)90155-w

Paul A Gagniuc. 2017. *Markov chains : from theory to implementation and experimentation.* John Wiley & Sons.

Hanna Halaburda, Zhiguo He, and Jiasun Li. 2021. An Economic Model of Consensus on Distributed Ledgers. https://www.nber.org/papers/w29515

Zhiguo He and Jiasun Li. 2022. Contract Enforcement and Decentralized Consensus: The Case of Slashing. https://ssrn.com/abstract=4036000

Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. 2001. In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. *American Economic Review* 91 (05 2001), 73–78. https://doi.org/10.1257/aer.91.2.73

Joseph Henrich and Michael Muthukrishna. 2020. The Origins and Psychology of Human Cooperation. 72 (2020). https://doi.org/10.1146/annurev-psych-081920-042106

Eric Horvitz. 2016. One hundred year study on artificial intelligence.

Sergey Ivanov. 2022. Reinforcement Learning Textbook. *arXiv preprint arXiv:2201.09746* (2022).

Nihal Koduri and Andrew W. Lo. 2021. The origin of cooperation. *Proceedings of the National Academy of Sciences* 118 (06 2021), e2015572118. https://doi.org/10.1073/pnas.2015572118

Ramakrishna Kotla, Allen Clement, Edmund Wong, Lorenzo Alvisi, and Mike Dahlin. 2007. Zyzzyva: Speculative Byzantine Fault Tolerance. https://www.cs.utexas.edu/users/dahlin/papers/Zyzzyva-CACM.pdf

Christian Kroer and Tuomas Sandholm. 2014. Extensive-form game abstraction with bounds. In *Proceedings of the fifteenth ACM conference on Economics and computation*. 621–638.

Leslie Lamport, Robert Shostak, and Marshall Pease. 1982. The Byzantine Generals Problem. *ACM Transactions on Programming Languages and Systems* 4 (07 1982), 382–401. https://doi.org/10.1145/357172.357176

Dan Levin and Luyao Zhang. 2020. Bridging Level-K to Nash Equilibrium. *The Review of Economics and Statistics* (10 2020), 1–44. https://doi.org/10.1162/rest_a_00990

Simon A. Levin and Andrew W. Lo. 2021. Introduction to PNAS special issue on evolutionary models of financial markets. *Proceedings of the National Academy of Sciences* 118 (06 2021). https://doi.org/10.1073/pnas.2104800118

Mohammad Hossein Manshaei, Murtuza Jadliwala, Anindya Maiti, and Mahdi Fooladgar. 2018. A Game-Theoretic Analysis of Shard-Based Permissionless Blockchains. *IEEE Access* 6 (2018), 78100–78112. https://doi.org/10.1109/access.2018.2884764

Tshilidzi Marwala and Bo Xing. 2018. Blockchain and Artificial Intelligence. *arXiv:1802.04451 [cs]* (10 2018). https://arxiv.org/abs/1802.04451

Eric Maskin and Jean Tirole. 2001a. Markov Perfect Equilibrium. *Journal of Economic Theory* 100 (10 2001), 191–219. https://doi.org/10.1006/jeth.2000.2785

Eric Maskin and Jean Tirole. 2001b. Markov perfect equilibrium: I. Observable actions. *Journal of Economic Theory* 100, 2 (2001), 191–219.

Alexander J McKenzie. 2009. Evolutionary Game Theory (Stanford Encyclopedia of Philosophy). https://plato.stanford.edu/entries/game-evolutionary/

Satoshi Nakamoto. 2008. Bitcoin: a Peer-to-Peer Electronic Cash System. https://bitcoin.org/bitcoin.pdf

Noam Nisan. 2007. *Algorithmic game theory*. Cambridge University Press.

Martin J Osborne and Ariel Rubinstein. 1994. *A course in game theory*. Mit Press.

José Parra Moyano and Omri Ross. 2017. KYC Optimization Using Distributed Ledger Technology. *Business & Information Systems Engineering* 59 (11 2017), 411–423. https://doi.org/10.1007/s12599-017-0504-2

Rafael Pass and Elaine Shi. 2017. Hybrid Consensus: Efficient Consensus in the Permissionless Model. In *DISC*.

M. Pease, R. Shostak, and L. Lamport. 1980. Reaching Agreement in the Presence of Faults. *J. ACM* 27 (04 1980), 228–234. https://doi.org/10.1145/322186.322188

Lee Ross, David Greene, and Pamela House. 1977. The “false consensus effect”: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology* 13 (05 1977), 279–301. https://doi.org/10.1016/0022-1031(77)90049-x

Ariel Rubinstein. 2002. *Modeling bounded rationality*. Mit Press.

Ariel Rubinstein and Yuval Salant. 2016. "Isn't everyone like me?": On the presence of self-similarity in strategic interactions. *Judgment and Decision Making* 11, 2 (2016), 168–173. https://EconPapers.repec.org/RePEc:jdm:journl:v:11:y:2016:i:2:p:168-173

Khaled Salah, M. Habib Ur Rehman, Nishara Nizamuddin, and Ala Al-Fuqaha. 2019. Blockchain for AI: Review and Open Research Challenges. *IEEE Access* 7 (2019), 10127–10149. https://doi.org/10.1109/access.2018.2890507

Fahad Saleh. 2018. Blockchain Without Waste: Proof-of-Stake. *SSRN Electronic Journal* (2018). https://doi.org/10.2139/ssrn.3183935

Fahad Saleh. 2021. Blockchain without waste: Proof-of-stake. *The Review of financial studies* 34, 3 (2021), 1156–1190.

Lakshmi Siva Sankar, M Sindhu, and M Sethumadhavan. 2017. Survey of consensus protocols on blockchain applications. In *2017 4th international conference on advanced computing and communication systems (ICACCS)*. IEEE, 1–5.

Frrddric Schneider and Roberto A. Weber. 2013. Long-Term Commitment and Cooperation. *SSRN Electronic Journal* (2013). https://doi.org/10.2139/ssrn.2334376

Sandip Sen. 2013. A comprehensive approach to trust management. (2013), 797–800. http://www.ifaamas.org/Proceedings/aamas2013/docs/p797.pdf

Robert Shimer and Lones Smith. 2000. Assortative Matching and Search. *Econometrica* 68 (03 2000), 343–369. https://doi.org/10.1111/1468-0262.00112

Yoav Shoham and Kevin Leyton-Brown. 2009. *Multiagent systems : algorithmic, game-theoretic, and logical foundations*. Cambridge University Press.

Herbert A. Simon. 1955. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics* 69 (1955), 99–118. https://doi.org/10.2307/1884852

JMPGR Smith and George R Price. 1973. The logic of animal conflict. *Nature* 246, 5427 (1973), 15–18.

John Maynard Smith. 1979. Game theory and the evolution of behaviour. *Proceedings of the Royal Society of London. Series B. Biological Sciences* 205, 1161 (1979), 475–488.

J. MAYNARD SMITH and G. R. PRICE. 1973. The Logic of Animal Conflict. *Nature* 246 (11 1973), 15–18. https://doi.org/10.1038/246015a0

John M. Staatz. 1983. The Cooperative as a Coalition: A Game-Theoretic Approach. *American Journal of Agricultural Economics* 65 (12 1983), 1084–1089. https://doi.org/10.2307/1240425

Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

Eva Tardos and Vijay V Vazirani. 2007. Basic solution concepts and computational issues. *Algorithmic game theory* (2007), 3–28.

Wei Tong, Xuewen Dong, and Jiawei Zheng. 2019. Trust-PBFT: A PeerTrust-Based Practical Byzantine Consensus Algorithm. , 344–349 pages. https://doi.org/10.1109/NaNA.2019.00066

TRON. 2018. Advanced Decentralized Blockchain Platform Whitepaper Version: 2.0 San Francisco. https://tron.network/static/doc/white_paper_v_2_0.pdf

Giuliana Santos Veronese, Miguel Correia, Alysson Neves Bessani, Lau Cheuk Lung, and Paulo Verissimo. 2013. Efficient Byzantine Fault-Tolerance. *IEEE Trans. Comput.* 62 (01 2013), 16–30. https://doi.org/10.1109/tc.2011.221

Feilong Wang, Yipeng Ji, Mingsheng Liu, Yangyang Li, Xiong Li, Xu Zhang, and Xiaojun Shi. 2021. An Optimization Strategy for PBFT Consensus Mechanism Based On Consortium Blockchain. *Proceedings of the 3rd ACM International Symposium on Blockchain and Secure Critical Infrastructure* (05 2021). https://doi.org/10.1145/3457337.3457843

Puming Wang, Laurence T Yang, Jintao Li, Jinjun Chen, and Shangqing Hu. 2019. Data fusion in cyber-physical-social systems: State-of-the-art and perspectives. *Information Fusion* 51 (2019), 42–57.

Yonghong Wang and Munindar P Singh. 2007. Formal trust model for multiagent systems. (2007), 1551–1556. https://www.aaai.org/Papers/IJCAI/2007/IJCAI07-250.pdf

Yang Xiao, Ning Zhang, Wenjing Lou, and Y Thomas Hou. 2020. A survey of distributed consensus protocols for blockchain networks. *IEEE Communications Surveys & Tutorials* 22, 2 (2020), 1432–1465.

Bo Xing and Tshilidzi Marwala. 2018. The Synergy of Blockchain and Artificial Intelligence. *SSRN Electronic Journal* (2018). https://doi.org/10.2139/ssrn.3225357

Chun-Lei Yang and Ching-Syang Jack Yue. 2019. Cooperation in an Assortative Matching Prisoners Dilemma Experiment with Pro-Social Dummies. *Scientific Reports* 9 (09 2019). https://doi.org/10.1038/s41598-019-50083-6

A GLOSSARY TABLE

B NOTATION TABLE

Table 7. Glossary Table

Glossary	Definition	Reference	Type
Practice Byzantine Fault Tolerance (pBFT)	A consensus algorithm to deal with the Byzantine General Problem.	Castro and Liskov [2002]	Consensus protocol
Proof-of-Work (PoW)	A form of cryptographic proof in which one party (the prover) proves to others (the verifiers) that a certain amount of a specific computational effort has been expended.	Nakamoto [2008]	Consensus protocol
Proof-of-Stake (PoS)	A class of consensus mechanisms for blockchains that work by selecting validators in proportion to their quantity of holdings in the associated cryptocurrency.	Saleh [2018]	Consensus protocol
Player/ Miner/ Agent / Node	The objects mining in blockchain. (These three words have the same representations in our paper)	[Castro and Liskov, 2002]	Distributed computing/ Game theory
Agreement	(1)The nonfaulty processors compute exactly the same vector. (2)The element of this vector corresponding to a given nonfaulty processor is the private value of that processor.	Pease et al. [1980]	Evaluation criteria
Validity	A decided value by any rational player is valid; it satisfies the predefined predicate.	Amoussou-Guenou et al. [2020a]	Evaluation criteria
Termination	Every rational player decides on a value (a block).	Amoussou-Guenou et al. [2020a]	Evaluation criteria
Consensus	At least 51% of the nodes on the network agree on the next global state of the network.		Evaluation criteria
Safety	All non-faulty replicas agree on the sequence numbers of requests that commit locally.	Castro and Liskov [2002]	Evaluation criteria
Liveness	Replicas must move to a new view if they are unable to execute a request	Castro and Liskov [2002]	Evaluation criteria
Strategy	A function that assigns an action to each nonterminal history	[Osborne and Rubinstein, 1994]	Game theory
Perfect Bayesian Equilibrium (PBE)	An equilibrium concept relevant for dynamic games with incomplete information (sequential Bayesian games)	[Fudenberg and Tirole, 1991c]	Game theory equilibrium solution concept
Markov Perfect Equilibrium (MPE)	A set of mixed strategies for each of the players that satisfies some criteria	[Maskin and Tirole, 2001a]	Game theory equilibrium solution concept

Table 8. Notation table

notation	definition
$\mathcal{A}$	a set $\{A_i\}_{i=1}^{N}$ with N elements, the committee of N miners established at time $t = 0$
A_i	an agent with order i
N	the number of agents in $\mathcal{A}$, the maximum value of i
t	the enumeration of game rounds
P_t	the selected proposer for round t
h_i	the proposal made by P_t at round t
R	the reward to the validators who send a message when the block is accepted
c_{check}	the cost to the validators who check the validity of the proposed block
c_{send}	the cost of the validators who send a message
κ	the cost occurs to all validators with the honest strategy when an invalid block is accepted
S_H	the honest strategy
S_B	the Byzantine strategy
s_i	the strategy chosen by agent A_i at round t, $s_i \in \{S_H, S_B\}$
ν	the majority threshold of the votes
x_t	the proportion of agents with S_H in round t
x_1	the initial proportion of agents with honest strategies
m	the portion of the rounds that a validator believes to meet a proposer with the same strategy in the game
$\pi_{ij}(x_t)$	the subjective meeting probability of one validator with strategy S_i meet one proposer with strategy S_j in function of x_t, where $i, j \in \{H, B\}$
$V_{ij}(x_t)$	the expected payoff of one validator with strategy S_i meet one proposer with strategy S_j in function of x_t, where $i, j \in \{H, B\}$
$V_i(x_t)$	the expected validator payoff with subjective belief of one validator with strategy S_i in function of x_t
P_{H_t}	the probability that an agent chooses S_H in round t
P_{B_t}	the probability that an agent chooses S_B in round t
α	the proportion of R to κ
β	the proportion of c_{check} to κ
γ	the proportion of ν to N